nature

www.nature.com/nature

Vol 450 | Issue no. 7172 | 13 December 2007

Directive action required

Europe's handling of applications to grow genetically modified crops amounts to bad governance.

It took many years of acrimonious debate for the European Union (EU) to agree a directive regulating the cultivation of genetically modified (GM) crops. In many member countries, the public was ready to accept genetic technologies in the service of medicine but not, as they saw it, in the service of the agricultural industry. That industry, aggressively in favour of GM crops, continues to be powerful and influential. European publics remain strongly opposed.

There was extensive consultation in formulating the directive, and science was recruited in support of each side. But six years after it was passed, not a single application has been approved for cultivation. Many EU countries are showing their continuing distaste for GM crops by refusing to grow the only one currently approved (authorized before the new rules came into effect), Monsanto's MON810 insect-resistant maize (corn). And last month, environment commissioner Stavros Dimas prepared to reject applications for two varieties of insect- and herbicide-resistant maize, from Syngenta and Pioneer Hi-Bred International, by inappropriately overturning the recommendation of his scientific advisers (see page 928).

This all highlights the problematic framework within the EU bodies. The approval process calls on the commission to make a science-based decision, gives member states the chance to decide politically on that science-based decision and then, if they can't agree, leaves the final decision entirely up to the commission. In the current cases, the commission is at war with itself, with powerful commissioners such as those for agriculture and industry trying to get Dimas to change his mind and recommend approval of the two crop varieties.

Dimas has misused science to tip the balance of his analysis of risks and benefits with which he justified his decision. Central to the process is the European Food Safety Authority (EFSA), which operates independently of the commission and the member states, and is mandated with securing independent scientific advice for them. Dimas is free to seek further, or even alternative, scientific advice. But his draft decision cites 11 papers purportedly demonstrating environmental risk that were published during the shamefully long period that the EFSA's report sat on his desk, without evaluating them in the context of the body of scientific literature. He has declined to respond to questions about how this selection of publications was made. This is neither in the letter nor the spirit of the directive. His decision to say 'no', where the EFSA said 'yes', is a political, not a scientific, move.

At this point, whatever the commission proposes in terms of these two crop applications, the member states are unlikely to be able to decide with the necessary majority one way or the other. And so the infighting commission will make the final decision.

The directive needs to be revised to ensure that the checks and balances put in place to reassure opponents, while not crushing innovation, cannot be abused by the political motives of one side. Most importantly, scientific input must be handled appropriately. So the proposal of German agriculture minister Horst Seehofer for a single agency dedicated to such tasks, which makes decisions based only on science and does not need to send every application for individual political approval, makes sense. This won't happen soon given the politics, but there is every reason to support the idea.

"Scientific input must be handled appropriately."

Scientists and others cannot reiterate too often that crops optimized to particular environments by genetic enhancement will be of significant benefit to societies in rich and poor countries, for example by increasing yields, allowing crops to grow in poorly fertile regions and reducing the amount of external chemical control required to maintain a healthy crop. The available evidence indicates that their potential for damaging the environment is small. Rigorous science-based risk assessment is likely to favour the cultivation of GM crops, subject to appropriate surveillance.

But whatever science indicates, member states want to protect their veto rights on crop applications because of opposition to GM crops by their publics. Benefits from the technology can be expected to become more apparent, for example in cheaper and better foods, or locally grown foods that help prevent famine. Until that time, advocates will need to persist against a strong political tide. Meanwhile, a directive that makes the fate of such crops dependent on the conflicted perspectives of individual commissioners is failing and needs repair. ■

Hollow victory

More benign global AIDS statistics do not mean that the battle against HIV is being won.

In a rare piece of good news, the Joint United Nations Programme on HIV/AIDS (UNAIDS) last month cut its estimate of the number of people infected with HIV worldwide. The revised figures brought the estimate of those infected down from 39.5 million to 33.2 million, and put the number of new infections for 2007 at 2.5 million, down from what the agency now says was a late-1990s peak of more than 3 million per year. The revised statistics are particularly encouraging for India, where the agency says that 2.5 million people are infected with HIV, a figure that is less than half its previous estimate.

But most of this change is accounted for by more accurate sampling techniques, and the new numbers sadly reflect little significant progress in combating AIDS on the ground. That is especially true in sub-Saharan Africa where UNAIDS reports that 68% of all HIV infections and 76% of AIDS deaths now occur. In these countries, women are often unable to insist on condom use, concurrent relationships contribute to the spread of the disease, and fewer than one-third

of patients that could benefit have access to the antiretroviral drugs that can considerably extend life expectancy.

Although immunologists and virologists are making progress in understanding, for example, how HIV affects the mucosal surfaces of the body — a process that seems to contribute greatly to HIV's destructive toll on the immune system — the search for a preventative vaccine has stalled. Only this autumn, a trial for a candidate vaccine from Merck was halted because it proved to be ineffective, and actually made some subjects more vulnerable to infection (see *Nature* **450,** 325; 2007). Some researchers are now claiming that disappointing lab results for this vaccine earlier in its development should have prevented it from getting to full-blown clinical trials. Those in the field have recently made attempts to ensure that new vaccine candidates meet rigorous scientific standards agreed to by the entire field — but this initiative began after the Merck vaccine had entered large clinical trials.

It seems that in vaccine development, researchers waited too long to coordinate their efforts fully. That could provide a lesson for the parallel quest for an effective microbicide — a chemical prevention method whose use, importantly, would be controlled by women. All the results of large microbicide trials to date have been disappointing — and duplicative microbicide trials are still being planned (see *Nature* **448,** 110–111; 2007). It is not clear that the most promising microbicide candidates are those that are being advanced most rapidly into trials, nor is there any consensus about what the most scientifically promising candidate would look like.

These are issues the microbicide field needs to resolve. With no vaccine in sight, the microbicide researchers are arguably those best placed to deliver something that will fundamentally alter the shape of the AIDS pandemic. That way, UNAIDS might one day deliver a downward estimate in the worldwide HIV burden that can be attributed to genuine progress against the disease, rather than to better statistical sampling. ■

Venezuela's way ahead

The opportunities currently opening up for Venezuelan science should not be squandered.

The president of Venezuela, Hugo Chávez, suffered his first electoral defeat for a decade on 2 December, when he unexpectedly lost a referendum on constitutional change that was supposed to cement his powers and accelerate socialist reform. The opposition was spearheaded by protest marches of hundreds of thousands of students, along with their professors. But the left-populist president, for all his flaws, has broadly supported universities and scientific research in Venezuela.

Chávez sees himself as the leader of a socialist revolution, modelled on the egalitarian ideals of Simón Bolívar, the Caracas-born general who led the liberation of much of South America from Spanish rule in the early nineteenth century. Chávez has nationalized major industries, including the oil companies, and has increasingly distanced Venezuela politically from the United States, its largest trading partner. Rapid economic growth has been sustained by the rising price of Venezuela's oil exports.

The Venezuelan president, while openly confronting the oil companies and other national élites, has taken steps to keep academics on his side. Like army officers, Venezuelan professors can retire at the age of 47 and receive generous pensions for the rest of their lives. Not everyone takes this up — but a sizeable fraction of the 33,000-strong academic workforce do just that. Professors also have the right to choose their own students. Their tendency to choose from the upper middle class may explain some of the student protests against Chávez's socialist government.

On the other hand, measures have been taken to strengthen the universities. In 2001, the government created a Ministry of Science and Technology, which distributes grant money on a competitive basis. And in January 2007 the Organic Law of Science, Technology and Innovation (LOCTI) came into effect, requiring Venezuela's 7,000 largest companies and commercial enterprises to pay a fraction of their annual taxes directly to universities and public research institutes. Overall public and private spending on science has quadrupled, to US$2.5 billion per year, the government says, reaching a very respectable 2.1% of gross domestic product in 2007.

As a result of these measures, some academics say, the Venezuelan science system is suddenly receiving more support than it can sensibly manage. Companies are investing in research projects as they see fit, without a proper system for evaluation of the proposed work. The government is now evaluating the first year of the work supported by LOCTI and must then find ways to channel more of the money into the most promising projects.

Obvious national research priorities range from infectious-disease research and rainforest ecology, to engineering and environmental problems related to oil retrieval. One problem is that few departments at Venezuela's 50 or so universities have sufficient staff and equipment to perform internationally competitive research. Another issue is that many professors are not especially interested in doing original research, as regular publication is not necessarily rewarded with promotion. Making research a prerequisite of a successful academic career — which should not end at the age of 47 — is the key to making Venezuelan science more productive.

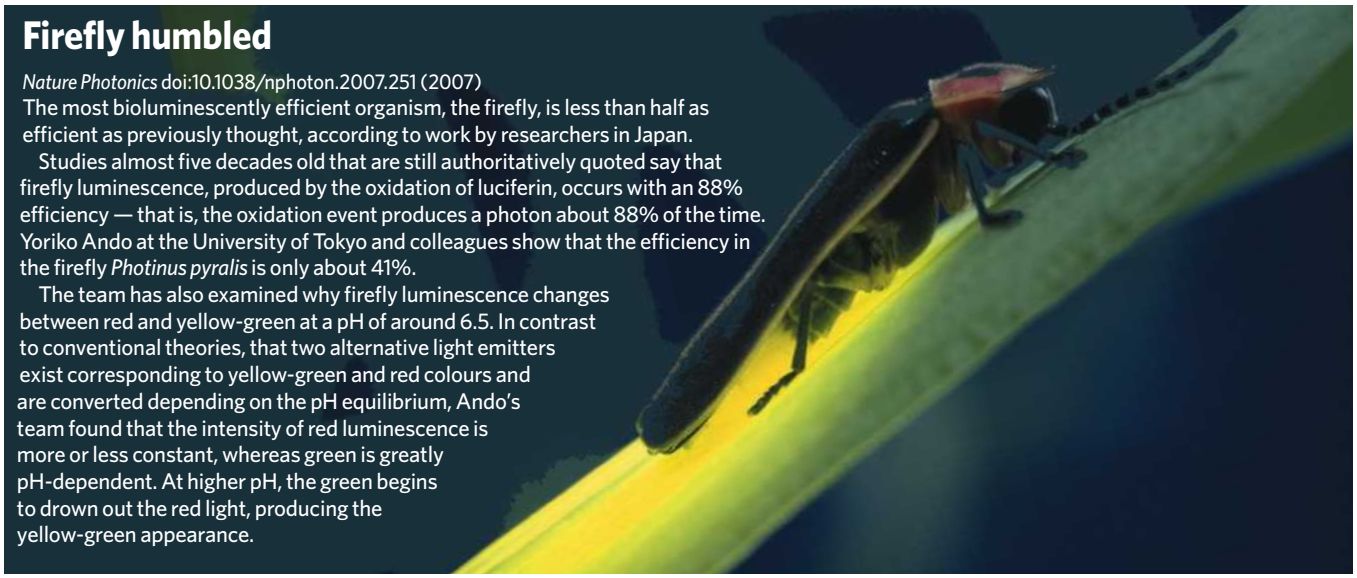
"Few departments at Venezuela's universities have sufficient staff and equipment to perform internationally competitive research."

Plans also exist to turn the country's premier research institute, the Venezuelan Institute for Scientific Research in Caracas, into a full-blown research university. This will help to produce qualified and motivated graduate students who can take Venezuelan science forward. The institute should have enough income from public and private sources to set up new centres in the Andes, the Amazon region and in the oil-rich state of Zulia in northwestern Venezuela — all of which need to raise their research profiles.

The referendum result has raised hopes that Venezuela's democracy will outlive Chávez, and build on some of his genuine achievements. The advent of stronger science at Venezuela's peripheries, as well as in its capital, is one legacy that could prove invaluable. ■

RESEARCH HIGHLIGHTS

Firefly humbled



RUNK/SCHOENBERGER/ALAMY

Nature Photonics doi:10.1038/nphoton.2007.251 (2007)

The most bioluminescently efficient organism, the firefly, is less than half as efficient as previously thought, according to work by researchers in Japan.

Studies almost five decades old that are still authoritatively quoted say that firefly luminescence, produced by the oxidation of luciferin, occurs with an 88% efficiency — that is, the oxidation event produces a photon about 88% of the time. Yoriko Ando at the University of Tokyo and colleagues show that the efficiency in the firefly *Photinus pyralis* is only about 41%.

The team has also examined why firefly luminescence changes between red and yellow-green at a pH of around 6.5. In contrast to conventional theories, that two alternative light emitters exist corresponding to yellow-green and red colours and are converted depending on the pH equilibrium, Ando's team found that the intensity of red luminescence is more or less constant, whereas green is greatly pH-dependent. At higher pH, the green begins to drown out the red light, producing the yellow-green appearance.

ANIMAL BEHAVIOUR

An elephant never forgets

Biol. Lett. doi:10.1098/rsbl.2007.0529 (2007)

Among the mobile elephant societies of Amboseli National Park in Kenya, the ability to keep track of kin may well be adaptive.

Richard Byrne of the University of St Andrews in Fife, UK, and his team tested whether African elephants (*Loxodonta africana*) could remember where specific individuals were. They did this by falsifying the presence of these individuals at unexpected locations using olfactory cues. The team gathered earth containing urine deposits and moved it to both likely and unlikely locations. In general, elephants reached for and sniffed urine more often and for longer if it was from elephants that were walking behind them or kin from far away, behaviour that the team interpreted as indications of surprise that the individual was in an unexpected place.

The differences between interest in urine in expected and unexpected locations were subtle, but the team guesses that each elephant keeps track of the locations of all the members of its group. Such groups can number as many as 30.

MATERIALS SCIENCE

Singled out

Phys. Rev. Lett. **99,** 227401 (2007)

By working in just one dimension, researchers have observed excitons — made up of an electron bound to a 'hole', which describes the absence of an electron — in a metallic system. Excitons are hugely important in semiconductor materials such as light-emitting devices and solar cells, but have never been observed in bulk metals because the charges of free electrons in the metal interfere with exciton formation.

Alex Zettl at the University of California, Berkeley, and his co-workers chose to hunt for excitons in metallic single-walled nanotubes, which are one-dimensional. The work confirms previous predictions that interference by free electrons would be reduced in one-dimensional conductors.

MEDICINE

Mirror mouse

J. Clin. Invest. doi:10.1172/JCI33284 (2007)

A genetic connection has been revealed between heterotaxy, in which the internal organs are positioned in a partial mirror image of their usual arrangement, and primary ciliary dyskinesia (PCD), in which cilia — tiny filamentous projections that extend from certain cell types — function poorly or not at all. Patients with heterotaxy often have heart defects, whereas those with PCD tend to have difficulty ridding their lungs of mucus and in males infertility is common, owing to immobile sperm.

Cecilia Lo, at the National Heart, Lung, and Blood Institute in Bethesda, Maryland, and her colleagues found a mutant mouse with a wrongly structured heart and a recessive mutation in a gene called *Dnahc5*. Patients with mutations in *DNAH5*, the human version of this gene, often have PCD. The researchers showed that 40% of mouse mutants with one copy of the recessive gene also exhibit heterotaxy. They conclude from breeding experiments that *Dnahc5* can cause both conditions, and suggest that many patients with heterotaxy may have undiagnosed PCD, and vice versa.

NANOTECHNOLOGY

Bitty barcodes

Nano Lett. 10.1021/nl072606s (2007)

Chemists have created nanowires that encode information in two ways. These could be used to detect disease or, after spraying them onto products or people, as tags for inventory management or espionage.

The nanowires act similarly to a barcode through the variable arrangement of pairs of gold disks spaced along each wire. The disks' paired arrangement, which includes a tiny space between them, amplifies a spectroscopic signal and allows it to be read. Chad Mirkin and his colleagues at Northwestern University in Evanston, Illinois, also added dyes that broadcast a unique spectrum when read by an instrument up to a third of a metre away.

To demonstrate the wires' use as biological detectors, the researchers fused single strands of DNA to the disks that, in solution, seek and bind their complementary target sequence. This could be used to test for diseases such as anthrax.

F. STÖBER/IMAGEBROKER/FLPA

GENETICS

The long and the short of it

Genes Dev. doi:10.1101/gad.1595107 (2007)

Short interfering RNAs (siRNAs) silence gene expression, regulating various cellular processes. Different types of naturally occurring siRNA exist, but the hallmark of these sequences is their short length of 20–31 nucleotides. Researchers now report that the model plant *Arabidopsis* also expresses long siRNAs (lsiRNAs) of 30–40 nucleotides.

Hailing Jin, from the University of California, Riverside, and her colleagues find that lsiRNAs share many features with other plant siRNAs, but differ in two aspects. Creation of lsiRNAs requires a unique set of proteins, and one lsiRNA seems to mediate the degradation of its target mRNA sequences by a mechanism that is unusual for plants. The lsiRNAs identified by the team are mainly induced in response to bacterial infection and under certain growth conditions.

ASTRONOMY

Galactic dust-busting

Astronom. J. **134,** 2385–2397 (2007)

Inside galaxies, stray dust blocks starlight and creates difficulties for astronomers. Most observations require a correction to compensate, but there have been relatively few measurements of dust in galaxies far from Earth.

Now Benne Holwerda of the Space Telescope Science Institute in Baltimore, Maryland, and his colleagues have used pairs of galaxies to gain a better understanding of far-away dust. Using the Sloan Digital Sky Survey, an archive of almost a quarter of the sky, the team selected 83 cases in which one galaxy was partially obscured by another (pictured, below). By comparing the exposed and obscured parts of the background galaxy, the team measured exactly how much light was absorbed by dust in the foreground galaxy.

The team looked at galaxies as far away as 2 billion light years, and is now working on extending the survey to even greater distances. The findings will aid a wide range of astronomical observations.

NASA/STSCI

CHEMISTRY

Special delivery

Nature Chem. Biol. doi:10.1038/nchembio.2007.56 (2007)

Salinosporamide A is a chlorinated natural product from a marine bacterium that is being tested in clinical trials for its cancer-beating properties. Bradley Moore, at the University of California, San Diego, and his colleagues have discovered an enzyme, SalL, that delivers chloride to a precursor of salinosporamide A by a unique enzymatic pathway.

Chlorine normally works its way into natural products by an oxidative mechanism. But in this case the researchers have identified a nucleophilic substitution that involves the breaking of a carbon–sulphur bond and the formation a carbon–chlorine bond. A biological methylating agent is highjacked by SalL to do this. Subsequent reactions of metabolite intermediates form salinosporamide A. The authors suggest that this enzyme opens up fresh possibilities for engineering metabolic pathways to create new chlorinated products.

NEUROBIOLOGY

Uncomfortably numb

Neuron **56,** 880–892 (2007)

During stressful situations, it pays to be able to cut out distractions. Now, researchers have found the mechanism that enables the stress hormone noradrenaline to block pain.

Pankaj Sah and his colleagues at the University of Queensland in Australia have found that in rats, noradrenaline prevents communication between a pain-sensing region of the brain called the parabrachial nucleus, and a portion of the central amygdala, a region that links emotion with sensory experience.

Axons from the parabrachial nucleus extend into the central amygdala. The researchers showed that noradrenaline acts on receptors in the parabrachial nucleus, leading to a reduction in the number of sites at which neurotransmitter is released into the central amygdala.

JOURNAL CLUB

Paul Mulvaney
University of Melbourne, Australia

A nanoscientist says block co-polymers may unblock nanotechnology.

One of the great drivers of the nanotechnology revolution has been the dream of molecular assembly. Essentially, this means using molecular or chemical forces to self-assemble smart or functional structures that could be integrated into electronic or optical devices.

Thomas Russell and his colleagues have recently provided a superb example of the way the field may be heading (B. Kim *et al. Small* **3,** 1869–1872; 2007).They took a diblock polymer — one that self-assembles into micelles — known as poly(styrene-b-4-vinylpyridine) and forced it to form microdomains by tuning the micelle structure through solvent exchange. This led to hexagonally ordered templates with periods of about 45 nanometres.

They then transferred these polymer templates by reactive ion etching to aluminium surfaces and fabricated regular pores by anodic oxidation at 4° C. The resulting hexagonal, close-packed pores are just 12 nanometres across, with nearest-neighbour spacings that should be tunable over a range of about 10–50 nanometres. The ordering extends over an area several micrometres square.

Particularly elegant is the seamless combination of colloid chemistry with more conventional 'top-down' processing. These wet chemical methods should be cheaper and more scalable than more conventional cleanroom-based techniques. What is particularly exciting about this approach is that the diblock structure can be readily tuned to provide a range of surface topologies and so a wide variety of potential templates. These could drastically simplify the fabrication of periodic, sub-wavelength structures for plasmonics-based applications such as optical circuitry.

Discuss this paper at **http://blogs.nature.com/nature/journalclub**

NEWS

London to host ambitious research hub

The announcement last week that Europe's largest medical-research facility is to be built in central London has been largely welcomed by Britain's biomedical community, which hopes that the centre will accelerate the translational research — so beloved by policy-makers — that brings discoveries from the lab to patients. But the process has ruffled a few scientists' feathers.

Prime Minister Gordon Brown says that the government will sell a plot of land between the British Library and the international train station at St Pancras to a consortium of the Medical Research Council (MRC), Cancer Research UK (CRUK), the Wellcome Trust and University College London, for £85 million (US$173 million). The total cost of the UK Centre for Medical Research and Innovation, including purchase of the land, is pegged at more than £500 million. The MRC and CRUK are expected to shoulder the bulk of the cost, and the Wellcome Trust has committed £100 million to the project. The centre, which is expected to open in 2013, will employ up to 1,500 researchers and support staff.

It aims to compete with other global multidisciplinary scientific-research collaborations such as Biopolis in Singapore, the Allston Initiative at Harvard University and the Science-based Zizhu Industrial Park in Shanghai. "Being in central London, right alongside main teaching hospitals and main offices for clinical research, is a much better location for translational research," MRC head Leszek Borysiewicz told *Nature*. "There is every opportunity for scientists to translate their work with the most appropriate clinical partners when they are that close to each other."

P. MACDIARMID/GETTY

International draw: the MRC's new labs will be next to St Pancras station.

Details of the research projects and teams that will be transferred to the centre remain scarce. Nobel laureate Paul Nurse, who is president of Rockefeller University in New York and CRUK's former director, will head an independent science-planning committee to determine the shape and direction of the centre's work and the facilities needed to carry it out.

The project will face numerous hurdles. Some researchers have expressed concerns that the infusion of funding into a prestigious project with limited space could ultimately hamper some basic research already taking place. And choosing to site the centre — which will include the largest animal-research laboratory in Europe and a category-4 virus containment laboratory — next to an international transport hub has sparked biosafety concerns.

Unease about potential staff reductions and shelving of core research has been particularly pronounced at the MRC's largest research body, the National Institute for Medical Research (NIMR), which will account for the bulk of the MRC's contribution to the new centre. The NIMR's 750 scientists and staff have already experienced nearly four years of debate within the MRC over its plans to move the institute from its current 19-hectare site in Mill Hill, northwest London, to central London. An earlier plan to move the institute to a site next to Euston Station in conjunction with University College London was ditched in March after the proposal faced scathing criticism from a key committee in the House of Commons that had held hearings on the plan.

NIMR staffers have previously expressed concerns that the 1.4-hectare site at St Pancras would have insufficient space for the institute's current facilities. Scientists have also criticized

Nuclear-reactor closure hits cancer tests

Hospitals across North America have been forced to cancel tests for cancer and heart disease because the unexpected closure of a Canadian nuclear reactor has led to a sudden shortage of medical isotopes.

The 50-year-old National Research Universal (NRU) reactor located in Chalk River, Ontario, was shut down on 18 November for scheduled maintenance and was due back online by mid-December. But Atomic Energy Canada, which owns and operates the facility, extended the outage to install safety-related equipment, including upgrades to the reactor cooling pumps. The reactor supplies about 60% of the molybdenum isotopes used in medical applications globally, including molybdenum-99, which decays into technetium-99m and is used in about 16 million nuclear medicine procedures annually in the United States.

"It's a disaster for patients," says Sandy McEwan, president of the Society of Nuclear Medicine. North American hospitals now have 20–30% of the medical isotopes they require, he says.

Hospitals use a generator to extract technetium-99m from a source of decaying molybdenum-99. A technetium-99m isotope has a useful life of about one week, but can be stretched to two. MDS Nordion, an Ottawa-based life-sciences firm and molybdenum supplier to Bristol-Myers Squibb Medical Imaging, says it expects shortages of the radioisotope until mid-January. Molybdenum-99 has a half-life of 66 hours and cannot be stockpiled. Reactors in Australia, South Africa and Brussels also produce molybdenum-99. The shortage has

MRC executives for poor internal communication over the course of the discussions on the institute's fate, and say that the continuing uncertainty over how much of the NIMR will be transferred to the new research centre is taking a toll on morale. "It's going to be a very difficult management process to keep people happy," says Robin Lovell-Badge, head of stem-cell biology and developmental genetics at the NIMR. "We are very nervous about it. Of course we can see the advantages and we want to be optimistic, but a lot of people at the institute just don't trust the MRC because of its past history."

Translational medicine

Borysiewicz says that Nurse's committee is expected to draw up broad outlines of the new centre's scientific mission over the weeks to come, and he adds that NIMR researchers will be represented on the committee. "It may be five or six years before the new site is ready," he says. "It is the MRC's intention to support the science at the NIMR during that period."

Others think that the move will help their work. Neil McDonald, a structural biologist at CRUK, notes that his current central London lab needs major refurbishment, and says that the new centre was being viewed positively by CRUK's several hundred researchers. "The advantage of a big research institute, and the synergies involved, are not just economies of scale but accessibility to facilities that you wouldn't be able to afford at a smaller institute," he says. "If you are in the same building and you see people in the canteen every day, it promotes collaborations and interactions, not just between research scientists but between scientists and clinicians."

The proposed centre still needs planning permission to go ahead, but with Brown's backing, it is likely to succeed. ■

Andrea Chipman

reignited a discussion over securing the US supply of medical isotopes by building a reactor in the United States.

The NRU reactor was to be decommissioned in 2005, but its operating licence was extended until problems with two replacement reactors — MAPLE 1 and 2 — could be solved. The two MAPLE reactors and a processing facility were designed to supply the entire global demand for molybdenum-99, iodine-131, iodine-125 and xenon-133. In June, Atomic Energy Canada said that it expected MAPLE 1 and the processing facility to be in service by October 2008, and MAPLE 2 by October 2009. ■

Hannah Hoag

CLIMATE CONFERENCE
Follow the UN meeting in Bali online.
www.nature.com/news/specials/bali/index

Enigmatic clouds illuminated

SAN FRANCISCO

New findings from the edge of space are unmasking Earth's highest clouds.

A NASA satellite called Aeronomy of Ice in the Mesosphere (AIM) is sending back the first detailed information on the 'noctilucent' clouds, which shimmer overhead just after sundown at high latitudes, where they reflect the below-horizon Sun.

The AIM data reveal that the clouds are ten times brighter than previously thought, and form at a broader range of altitudes. Although it had been thought that the clouds were limited to a single altitude band of 82 kilometres, in fact they can form at anywhere between 79 and 90 kilometres. AIM has also shown that small patchy groupings of cloud can grow dramatically with even a tiny fall in temperature.

"This all yields critical information to determine why these clouds form and vary," says atmospheric scientist James Russell of Hampton University in Virginia, AIM's principal investigator.

Noctilucent clouds form when water vapour condenses onto 'seed' particles in the mesosphere, the layer of the atmosphere that extends from about 50 to 80 kilometres up. They have appeared more frequently and at lower latitudes in recent years, perhaps as a result of rising concentrations of greenhouse gases.

"This all yields critical information to determine why these clouds form and vary."

Increasing amounts of methane can result in more water vapour at the relevant altitudes, and rising levels of carbon dioxide cause temperatures in the mesosphere to drop, enhancing the conditions in which the clouds can occur. Noctilucent clouds typically form at temperatures of between −134 and −148 °C.

The new details from AIM, reported on 10 December at a meeting of the American Geophysical Union in San Francisco, California, include strange 'ice rings' that appear in some of the clouds, says project co-investigator Gary Thomas of the University of Colorado in Boulder.

These crescent-shaped blobs may be caused by atmospheric disturbances that propagate upwards from near Earth's surface — a phenomenon not seen before. "If true, it opens up an entirely new mechanism we had not thought of before this mission was launched," says Russell.

AIM's photographs are far more detailed than earlier studies, with a resolution of 5 kilometres. "The detail is so much richer with AIM," says Matt DeLand of Science Systems and Applications, Inc. in Lanham, Maryland, who is not involved with the mission but works on other satellite data. "It's fascinating stuff."

The first AIM data came from noctilucent clouds over the Arctic. The mission is currently gathering data on clouds in the Southern Hemisphere. ■

Alexandra Witze

P. PARVIAINEN/SPL

Cloudscape: noctilucent clouds are seen in twilight skies at high latitudes.

SPECIAL REPORT

Showdown for Europe

The European Union is set to make a landmark decision on genetically modified crops, as **Alison Abbott** and **Quirin Schiermeier** report.

A mammoth bureaucratic battle is looming between senior European Commission officials and national governments that could affect the long-term prospects for the cultivation of genetically modified crops on the continent.

Late last month, the European Commission's environment commissioner Stavros Dimas said that he plans to reject applications from Syngenta and Pioneer Hi-Bred International for approval to grow two transgenic strains of maize (corn), on the grounds that the crops could adversely affect the environment.

Dimas's position has been welcomed by environmental groups and attacked by industry lobbyists. And researchers point out that it ignores the recommendation Dimas received from his own scientific advisers.

But the environment commissioner's move is far from the end of the matter. Behind-the-scenes battles are under way inside the commission, where a powerful faction wants Europe to accept genetically modified crops. That would avoid further conflict with the United States, which has complained to the World Trade Organization (WTO) that European reluctance to approve the crops amounts to protectionism.

In particular, the commissioners who are respectively responsible for trade, industry and agriculture — Peter Mandelson, Günter Verheugen and Mariann Fischer Boel — are trying to overturn Dimas's decision.

Observers on both sides of the debate say that, when the dust settles, it is quite possible the European Commission will give the green light to Syngenta's Bt11 maize and Pioneer's 1507 maize, which are genetically engineered to be resistant to both pests and herbicides.

At present, only one transgenic crop can be cultivated in Europe: Monsanto's MON810 insect-resistant maize, which now comprises nearly 2% of maize grown in Europe, most of it in Spain and France (see 'Transgenic maize'). MON810 was approved before 2001, when the European Union (EU) agreed a directive setting out complex rules for the future approval of such crops.

Getting the directive agreed in the first place took several years, and came with the proviso that there would be no approvals for import or cultivation until water-tight mechanisms for tracing the transgene, labelling transgenic seeds and governing the safe 'coexistence' of conventional and transgenic plants were in place. That took until 2004. Even as they voted for the directive, some countries — Austria, Luxembourg, Greece, France, Denmark and Italy — made it clear that they were still reluctanct to allow the crops in, arguing that the directive should have explicitly taken into account public opinion, which they say is firmly opposed to their cultivation.

Under the directive, each candidate strain is assessed for its impact on animal and human health and the environment before a decision is made on whether to approve its cultivation.

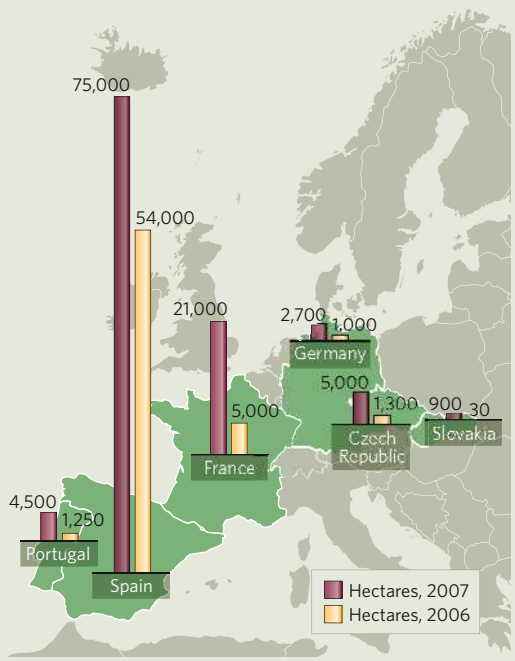
If a company wants to grow or market a crop in Europe (as food, feed or a derived product), it must apply through a member state. That country can either perform a scientific risk assessment itself for the commission or pass the application to the European Food Safety Authority (EFSA) in Parma, Italy, which organizes an assessment through a panel of 21 outside scientists. The EFSA delivers a scientific opinion to the commission's health directorate within six months. Five applications for the import of transgenic maize and oil-seed rape have been approved by this route since 2004.

But if the application includes cultivation of the crop, a more extensive environmental risk analysis must be carried out, and this is incorporated into the final scientific opinion delivered to the commission's directorate for the environment.

The commission should make a decision within three months. And this is the point where Bt11 and 1507 maize have got stuck.

Genetically modified maize makes up almost 2% of the crop grown in Europe.

TRANSGENIC MAIZE



Monsanto's MON810 maize, which is resistant to the European corn borer, is the only genetically modified crop that can be grown in the European Union (EU). It was approved in 1998, and this year varieties derived from it covered some 110,000 hectares of EU farmland across six countries — less than 2% of the total EU maize cultivation area.

Although cultivation is relatively high in Spain, Europe's use of transgenic maize is low compared with that of the United States, where 27.4 million hectares — three-quarters of the entire US maize crop — were grown this year. Around 38% of transgenic maize grown by US farmers already comes from seeds with combined resistance to pests and herbicides, such as the Bt11 and 1507 varieties for which approval in the EU is currently being sought.

Global cultivation of transgenic crops exceeded 100 million hectares for the first time in 2006. More than 10 million farmers in 22 countries use transgenic seed to grow soy, maize, oil-seed rape, cotton, rice and vegetables. **Q.S.**

C. KNAPTON/SPL

EFSA scientific reports on both varieties concluded that neither would have "an adverse effect on human and animal health or the environment" in the contexts proposed. Both reports were ready by April 2005, and were updated in November 2006.

But it wasn't until last month that a draft decision was circulated inside the European Commission saying that neither crop should be approved for cultivation. It refers to 11 papers published since the EFSA's update that it says cast doubt on the crops' long-term environmental safety.

The publications include studies claiming that insecticidal molecules from the plant may persist in water or sediments draining from a cultivated field, and may disturb downstream ecosystems.

The environment commissioner did not ask the EFSA panel for an opinion on these additional papers. Garlich von Essen, secretary-general of the European Seed Association, says that this shows "disdain" for both the EFSA and its advisory system.

Marc Van Montagu, a plant geneticist and president of the European Federation of Biotechnology, says the commission has cherry-picked publications claiming possible dangers, and he questions the quality of the selected papers. Environment-commission officials respond that their risk-management process is supposed to reach beyond the EFSA's findings.

Once the commission's decision has been finalized, it will go to the Standing Committee on the Food Chain and Animal Health, which comprises scientists and officials from member states.

The standing committee will vote on each proposal using a system — called the qualified majority vote — that reflects the size and population of each member state. If the voting is at odds with the commission's position, the dossiers are passed to the EU Environment Council of environment ministers of each member state, who must also vote on each case. But with populous nations such as Spain and the United Kingdom supporting approval, and Poland, Hungary and the Czech Republic joining some of the original dissenters, neither side is likely to obtain the two-thirds majority needed to decide the issue. If that happens, under EU rules the final decision will be thrown back to the commission itself.

On 26 November, the German agriculture minister Horst Seehofer proposed that this tortuous approval process should be abandoned and a regulatory authority be created with full responsibility for analysing the science and drawing conclusions.

"The reservations of the public are not being sufficiently considered," he said. "Until such an authority is created, there should be a moratorium on granting new approvals".

Such reservations are exemplified by the continued resistance of some nations to cultivating MON810 maize, which is grown in only six EU countries. Austria actually banned the import of the maize in 1997, and has since resisted strong pressure to lift the ban, which is illegal under the 2001 directive.

J. THYS/AFP/GETTY

Environment commissioner Stavros Dimas plans to reject applications to cultivate two transgenic crops.

In October, France's president Nicolas Sarkozy announced a suspension of the cultivation of transgenic maize until new national rules have been worked out. Sarkozy, who has recently laid out plans for far-reaching environmental improvements in France, seems willing to risk a dispute with the commission (and the WTO) over the issue.

Meanwhile, the WTO is putting increasing pressure on the EU, giving it until 11 January to end national moratoriums. The commission says it expects to make its decision on the two maize varieties in January as well, although an exact date has not been set. "This is a real test case," says Adrian Bebb, a Brussels-based campaigner for Friends of the Earth. "But we fear that Dimas's chances of winning are slim." ■

See Editorial, page 921.

ON THE RECORD

"I would have thought an intelligent person would have at least kept quiet until after tenure. Then you could advocate blowing up the Moon."

Bruce Harmon, a physicist at Iowa State University, muses in a recently released e-mail about astronomer Guillermo Gonzalez, who promoted intelligent design while seeking tenure. Gonzalez was turned down in May and is now appealing against the decision on the grounds of discrimination.

SCORECARD

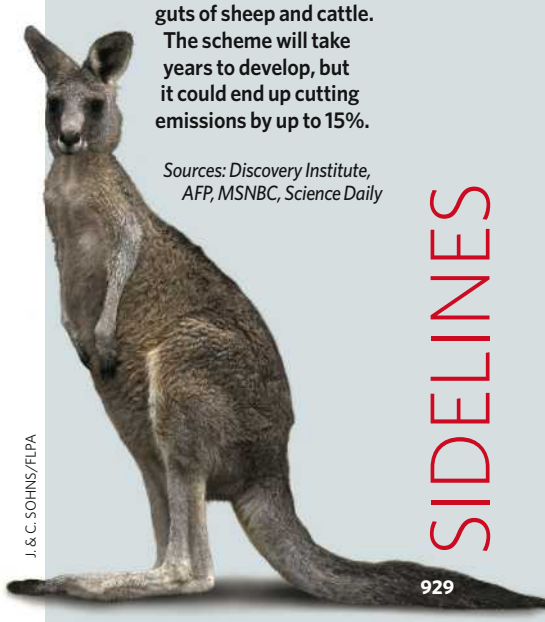
Coach potatoes
A year-long study finds that automated phone calls encouraging sedentary adults to exercise actually work.

Moon shots
Accusations that China faked photos from its Chang'e-1 Moon probe have proved to be false, but the investigation revealed flaws in the way the composite image was assembled, leaving the nation's space agency red-faced.

ZOO NEWS

Kangaroo flatulence
Kangaroos' farts are environmentally friendly. The marsupials' stomachs are home to bacteria that don't produce methane — a major greenhouse gas. This week Australian researchers unveiled a plan to transfer the green bugs into the guts of sheep and cattle. The scheme will take years to develop, but it could end up cutting emissions by up to 15%.

Sources: Discovery Institute, AFP, MSNBC, Science Daily



J. & C. SOHNS/FLPA

SIDELINES

Moonlighting missions

The next big NASA flagship mission to the outer planets will go to one of the moons of Jupiter or Saturn. By 1 January, the agency plans to narrow down the four candidate missions to two or three. It wants to keep costs below US$3 billion, but is planning to partner with the European Space Agency, which could contribute almost $1 billion. The final decision should be made by the end of 2008. **Eric Hand** weighs up their chances.

	EUROPA EXPLORER (JUPITER)	JUPITER SYSTEMS ORBITER / GANYMEDE	ENCELADUS EXPLORER (SATURN)	TITAN EXPLORER WITH ORBITER (SATURN)
Cost	$3.3 billion	$3.1 billion	$2.1 billion–$2.4 billion	$4 billion
Orbiter destination	A year spent circling above Europa, which has a saltwater ocean underneath a thin cap of ice.	A three-year tour of the jovian system, with many fly-bys of Europa, Io and Callisto, before settling into orbit around Ganymede.	To Enceladus, a frozen ball of ice only 500 kilometres in diameter. The Cassini mission discovered a geyser there spewing a plume of water into space.	It would use Titan's thick atmosphere to slow the spacecraft into orbit. This 'aerocapture' process reduces the amount of fuel needed and would allow a hot-air balloon and a lander to be dropped.
The pay-off	Data obtained by ice-penetrating radar would end a debate about the thickness of the ice shell and how often the water reaches the surface. Spectroscopy could detect organic signatures in any recent watery outbursts on the surface.	Scientists want to understand why Ganymede is the only moon in the Solar System with its own magnetic field.	The orbiter would pass through the water plume a dozen times, offering a rare chance to evaluate the biological potential of sub-surface water without landing or complicated drilling.	Like Earth, Titan has a 'water' cycle including rain and lakes — but with methane rather than water. Water erupting from ice volcanoes may combine with organic compounds to make amino acids — but scientists need a lander to test for them.
The challenge	Protecting the orbiter from Jupiter's radiation is expensive. Another cost is assembling the probe in a sterile environment so that microbes from Earth don't mix with those from Europa when the probe crashes to the surface.	The planet's radiation. The orbiter would have to survive for five years in Jupiter's harsh environment.	Enceladus is tiny and sits close to Saturn, which means it takes a lot of fuel to brake into its orbit. The geyser might only be spewing sublimated ice, rather than water from a sub-surface ocean. And what if the geyser stops before the probe gets there?	The Huygens probe's batteries lasted for just 90 minutes. A year-long balloon and lander mission would require expensive radioisotope power. Aerocapture hasn't been tested outside Earth.
Planned launch	2015	2017	2018	2018
The tip-book	**The one to beat.** Europa has been a top priority in NASA road maps and surveys. But some wonder whether it is worth visiting without actually landing there and sampling the ice and ocean directly.	**Dark horse.** The Jupiter community got a major mission 20 years ago with Galileo. And Juno, a smaller Jupiter mission, is scheduled for a 2011 launch. But proponents say that the whole-system approach gets the most bang for the buck.	**Wait until next time.** Enceladus wouldn't have even been considered but for the discovery of the plume. There are too many risks for an uncertain gain.	**Up and coming.** A Titan mission probably offers the most opportunities for ground-breaking science — and the only chance for an affordable lander, not to mention a balloon. But it will be hard to justify going to the same place as Huygens.

*Moons not to scale

DLR; JPL; JPL; JPL; NASA; H.RICHER (UNIV. BRITISH COLUMBIA)

Poor nations claim victory at climate talks

Negotiators at climate talks in Bali on Monday agreed on the details of a fund to help the most vulnerable nations adapt to climate change.

The adaptation fund, generated through a 2% levy on transactions between parties participating in the Kyoto Protocol's emissions-trading mechanism, has been accumulating for several years. But it has not yet been activated because countries could not agree on which body would approve the projects to be included in the scheme. They have now settled on the United Nation's Framework Convention on Climate Change.

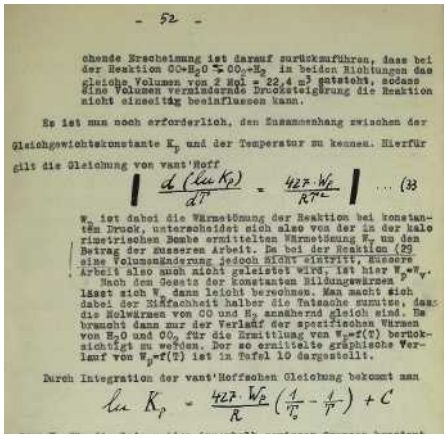
The decision is a victory for poor countries and small island states, which had argued that the fund, currently administered by a multilateral environmental-funding agency based in Washington DC, should have its own board and governance system. So far, the emissions-trading mechanism has generated US$67 million for adaptation measures.

Meanwhile, scientists in Bali called on governments to stabilize global emissions well below 450 parts per million of carbon-dioxide equivalents in the long term, their most concrete recommendation to politicians so far. Higher concentrations would mean that climate change may become unmanageable, they warned in a formal declaration issued on 6 December.

'Heated bidding' for rocket scientist's thesis

A formerly top-secret document produced by controversial rocket scientist Wernher von Braun for his PhD dissertation sold for US$33,000 at auction in New York last week.

The 166-page thesis, *Design, Theoretical and Experimental Contributions to the Problem of the Liquid-Fuel Rocket*,



- 52 -

chende Erscheinung ist darauf zurückzuführen, dass bei der Reaktion $CO+H_2O \rightleftharpoons CO_2+H_2$ in beiden Richtungen das gleiche Volumen von 2 Mol = 22,4 m³ gebildet, sodass ein Volumen vermindernde Druckerhöhung die Reaktion nicht einseitig beeinflussen kann.

Es ist nun noch erforderlich, den Zusammenhang zwischen der Gleichgewichtskonstante K_p und der Temperatur zu kennen. Hierfür gilt die Gleichung von vant'Hoff

$$\frac{d(\ln K_p)}{dT} = \frac{427 \cdot W_p}{RT^2} \quad \ldots (33$$

W_p ist dabei die Wärmetönung der Reaktion bei konstantem Druck, unterscheidet sich also von der in der kalorimetrischen Bombe ermittelten Wärmetönung W_v um den Betrag der äusseren Arbeit. Da bei der Reaktion (29) eine Volumenänderung jedoch nicht eintritt, äussere Arbeit also auch nicht geleistet wird, ist hier $W_p = W_v$.

Nach dem Gesetz der konstanten Bildungswärmen lässt sich W_v dann leicht berechnen. Man macht sich dabei der Einfachheit halber die Tatsache zunutze, dass die Bildungswärmen von CO und H_2 annähernd gleich sind. Es braucht dann nur der Verlauf der spezifischen Wärmen von H_2O und CO_2 für die Ermittlung von $W_v = f(T)$ berücksichtigt zu werden. Der so ermittelte graphische Verlauf von $W_p = f(T)$ ist in Tafel 10 dargestellt.

Durch Integration der vant'Hoffschen Gleichung bekommt man

$$\ln K_p = \frac{427 \cdot W_p}{R}\left(\frac{1}{T_0} - \frac{1}{T}\right) + C$$

Wernher von Braun's PhD thesis sold for $33,000.

was classified as 'top secret' by the Nazi government and, although written in 1934, was not published until 1960.

After working on rockets for the Nazis during the Second World War — his exact motivations are still debated — von Braun was spirited to the United States to work on missiles. He later became the first director of NASA's Marshall Space Flight Center in Huntsville, Alabama.

After what the auction house Bonhams describes as "heated bidding", the thesis sold for $3,000 more than its original estimate.

BONHAMS

Switzerland launches systems-biology initiative

One of the world's largest systems-biology research consortiums, known as SystemsX, has been launched in Switzerland.

The federal government has committed 200 million Swiss francs (US$177 million) to the project, half of which will support technology platforms at an ETH Zurich department based in Basel, which is also home to several large drug and chemical firms including Novartis and Roche.

The rest of the money will fund researchers at eleven universities and research institutes — or industry bodies — which will have to match those funds with their own money.

The Swiss National Science Foundation will supervise the scientific quality of the initiative, with help from international experts. This is a first for the granting agency, which has not previously been involved in the quality control of projects it does not fund itself.

US politicians push for food-safety funding boost

Twenty-three US senators are calling on President George W. Bush to boost funding for food-safety oversight in 2009. In a 6 December letter to Bush, the bipartisan group complained that the budget of the Food and Drug Administration (FDA) does not reflect its "critical" and growing role.

They noted, for instance, that the value of US agricultural imports had grown by 40% between 2003 and 2006, yet between 2004 and 2007, the number of employees dedicated to food safety at the FDA fell by 15% to 2,613. In February, Bush proposed increasing the agency's food-safety budget by $10.5 million, to $467 million. Congress has yet to approve the spending bill.

The letter comes on the heels of a highly critical report on 29 November from the FDA's scientific advisory board. It said that the $1.9-billion agency cannot fulfil its mission because of the erosion and inadequacy of its scientific base and information-technology infrastucture.

Private funds raise hopes for giant telescope

Plans to build the world's largest optical telescope were jump-started with a 5 December announcement that a foundation set up by Intel co-founder Gordon Moore and his wife Betty had given the California Institute of Technology and the University of California $200 million.

The two universities will also put in $100 million for the billion-dollar project, called the Thirty Meter Telescope (TMT).

The gift puts the TMT ahead of two other planned mega-telescopes — the Giant Magellan Telescope, a 24.5-metre telescope led by a consortium including the Carnegie Institution of Washington, and the Extremely Large Telescope, a 42-metre observatory planned by the European Southern Observatory. The TMT's 492 hexagonal mirrors will stretch for 30 metres, and the device is expected to achieve a better resolution than that of the Hubble Space Telescope. A final design is expected in 2009.

THIRTY METER TELESCOPE

The Thirty Meter Telescope will have 492 mirrors.

Hackers steal personal data from US laboratories

The Oak Ridge National Laboratory (ORNL) in Tennessee has warned some 12,000 people that their personal data may have been stolen as part of a "sophisticated cyber attack".

Hackers sent lab employees e-mails that seemed legitimate but contained attachments that, when opened, gave the hackers access to their computers. The 'phishing' scheme apparently allowed the hackers to download personal information about people visiting the laboratory between 1990 and 2004.

Correction

The legend to our graphic 'CO_2 emissions 1990–2006' (*Nature* **447,** 1038; 2007) erroneously gave CO_2 intensity in tonnes per thousand US$ GDP. It should have been tonnes per million US$ GDP.

BUSINESS

Promise boiling over

Geothermal power is one of the hottest prospects in the burgeoning clean-energy market. But, as **Kurt Kleiner** reports, it's not close enough to home for many uses.

Iceland is famously rich in geothermal energy. The country sits on a geological hot spot that provides enough power to generate one-quarter of its electricity and heat 90% of its homes.

Now, the Icelandic bank Glitnir has decided that the time is ripe to take advantage of geothermal opportunities elsewhere. In September, the bank opened an office in New York to pursue what it boldly predicts will be $40 billion worth of geothermal investment in the United States over the next 20 or so years.

Glitnir's move is one of a growing number of signs that geothermal energy is ready to become a more significant player in world energy production. "If you're a utility, your first choice of renewable energy is geothermal," says Thomas King, managing director of the US Renewables Group investment fund in New York. "It's the cream of the crop."

King's bullishness reflects a growing belief among energy analysts that although the technology hasn't received as much attention as wave or solar power, geothermal companies have outstanding long-term potential. Robert Wilder, chief executive of Californian clean-energy consultancy WilderShares, points to Ormat Technologies, a maker of geothermal plants based in Reno, Nevada, as a sign of the trend: its share price has risen from about $16 a share in April 2005 to $50 this week.

Nevertheless, with just 9 gigawatts or so of installed capacity, geothermal energy accounts for only about 0.2% of all electricity produced around the world. In theory, geothermal heat can be found anywhere in the world if you dig deep enough. But in practice, it has only been worth harnessing in regions where water is found in combination with hot, porous rock close to the surface.

"A well-managed reservoir can keep going practically forever."

For instance, just north of San Francisco, a geothermal field called The Geysers generates 760 megawatts of electric power. The plants there take advantage of a large, naturally occurring underground steam reservoir that can be tapped by drilling relatively shallow wells.

The Geysers are examples of 'dry-steam' power plants: the steam that comes out of the reservoir contains little or no liquid water, and can therefore be routed directly to a turbine to create electricity. However, most plants are of the 'flash-steam' variety. These plants use water that has been heated to about 180 °C, but remains liquid because it is highly pressurized underground. The water is then pumped to the surface. Because the pressure there is lower, most of the water 'flashes' into steam, which can be used to operate a turbine.

When the water is around 100–180 °C, 'binary' plants are used. These make use of hot water to heat a working fluid, such as isobutane, which has a lower boiling point than water does. The vaporized working fluid is then used to drive the turbine.

Deep pockets

Even in favourable geological locations, geothermal power has a high capital cost, mainly because it costs a lot to dig the wells. Balancing that are its low fuel costs. Overall, conventional geothermal plants in the United States deliver electricity for between 5 cents and 8 cents per kilowatt-hour: not much more than the average of 4 cents per kilowatt-hour for electricity from a coal-fired power plant.

In Australia, a firm called Geodynamics is trying to develop a new technique that can take advantage of geothermal energy in the absence of a ready-formed reservoir of water.

"We believe that our area is probably the best location in the world to make this approach economically viable," says Doone Wyborn, a founder and executive director of Geodynamics in Queensland.

The firm plans to use the Cooper Basin, a geological feature of the Australian interior, in which rocks with a temperature of about 270 °C are available quite close to the surface.

Geodynamics aims to drill two wells to a depth of 4.3 kilometres, and to fracture the hot granite in the rocks by pumping down cold water. Once the rock is permeable enough, the system will act as a heat exchanger — water will be pumped down one well, migrate through the rock to the other well, from which it will be extracted and used to generate electricity.

The company plans to have a 50-megawatt power plant in operation by 2010. It estimates the potential capacity of the Cooper Basin at 10,000 megawatts of power, which could be realized by drilling hundreds of wells.

CALPINE

Power plants from California (left) to Iceland tap into Earth's vast supplies of geothermal energy.

The Geodynamics project is an example of a technology called 'hot-dry-rock' or 'hot-fractured-rock' geothermal. A report by the Massachusetts Institute of Technology (MIT) in Cambridge published in January concluded that this type of 'enhanced' geothermal power generation could greatly enhance our ability to tap geothermal energy. "I feel it's been an ignored option," says Jefferson Tester, the chemical engineer at MIT who headed the panel that wrote the report, *The Future of Geothermal Energy*. "But I'm very optimistic about the possibility if a lot of things come into place."

Eventually, geothermal energy could be available almost everywhere, the report contends. Deep drilling from any location will eventually hit hot rock. In the United States alone, the report says, the amount of energy available by drilling up to 10 kilometres below the surface is a stunning 13 yottajoules (10^{24} joules), or 130,000 times the annual energy consumption of the entire country.

Only a fraction of that is economical to

O. POPOV/REUTERS

exploit. Even so, the report concluded, in the United States alone, enhanced geothermal electrical capacity could reach 100 gigawatts in the next 50 years — enough to fill about 10% of the country's electricity needs.

An important benefit of such systems is their flexibility, Tester says. They could prove to be economical from a very large scale, all the way down to a relatively small, 1-megawatt plant that also provides direct heating to buildings. As in Iceland, this combined heat-and-power approach greatly enhances the economics of geothermal power. But it requires building communities that can make use of the heat.

Powerful exchanges

Another geothermal option is the use of heat pumps. These use a vapour-compression cycle — the same principle that makes a refrigerator work — to transfer heat from below the ground in the winter, and to transfer excess heat out from buildings in the summer.

Tester says that geothermal power will make economic sense even without special incentives or restrictions on carbon emissions. As governments move to restrict greenhouse-gas emissions, geothermal power is set to look even better.

King adds that the standards for renewable energy being set by individual states have kicked off a flurry of interest in geothermal power. "It's clean, it's close to zero emissions, and it's baseload power that runs 24 hours a day, 7 days a week. And a well-managed reservoir can keep going practically forever," he says. ■

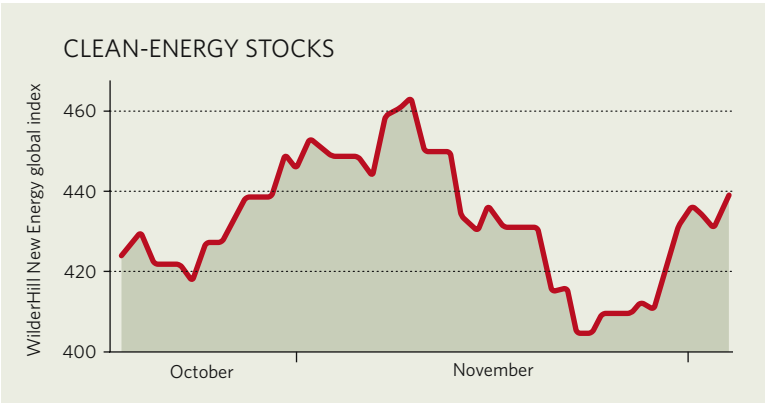
IN BRIEF

EURO TIE-UP Swiss drug giant Novartis has extended its partnership with MorphoSys, a German biotechnology firm, saying it will pay at least US$600 million over 10 years as the price of working with the smaller company to help it discover and develop antibody-based medicines. If Munich-based MorphoSys meets milestones for clinical development and market approval, the payments could exceed $1 billion, the companies said on 2 December. Novartis will gain almost exclusive access to the smaller firm's libraries of human antibodies, copies of which will be transferred to Novartis research sites.

MUTUAL MOUSE Sanofi-Aventis has struck a deal with US biotech firm Regeneron Pharmaceuticals to discover and develop therapies with a genetically engineered 'super-mouse' that can make human antibodies. Under a deal announced on 29 November, the Paris-based drugmaker will pay Regeneron, based in Tarrytown, New York, $85 million up front, and up to $475 million in research and development costs over the next 5 years. Sanofi will buy $312 million of stock in the smaller firm, boosting its stake from 4% to 19%.

NO APPROVAL An advisory panel to the US Food and Drug Administration has ruled that the cancer drug Avastin (bevacizumab) should not be approved for the treatment of advanced breast cancer. Shares in Genentech, the California-based biotechnology company that makes the drug, fell by $10 to $66 as news of the decision emerged last week. European regulators approved the drug for breast cancer use earlier in the year, but the FDA panel heard that although the drug helped some patients there was insufficient evidence that it improved their long-term survival prospects, and voted 5–4 against its approval.

MARKET WATCH



The inexorable rise of global renewable energy stocks — as measured by the WilderHill New Energy Global Innovation Index (symbol NEX on the US stock exchange) juddered to a halt last month.

Specialists attribute the reverse to doubts about pending US energy legislation as well as to the global credit crunch, which has hit stock markets in general.

The US Congress is currently considering an energy bill that will provide tax credits for renewable energy. But growing doubts about the extent of the subsidies have worried investors who had hoped that the bill would spur demand for wind and solar-power equipment, says Robert Wilder, founder of WilderShares, a Californian consultancy that co-compiles the index with New Energy Finance of London. He adds that biofuels stocks have "crashed and burned" as doubts grow about the commercial usefulness of the technology.

However it is concern about the availability of credit in the economy that has hit shares hardest, says Michael Leibreich, founder of New Energy Finance. He says that the market has been on "a bit of a roller-coaster ride", but notes that when all is said and done, it is still up 50% since the start of the year. "This reflects the extraordinary level of interest in the sector, evidenced by the continuing wave of money washing into it," he says.

A huge initial public offering planned for 13 December by Iberdrola Renewables, the renewables arm of Madrid-based utility Iberdrola, could raise up to €6 billion (US$8.8 billion), reminding the markets of this strength, Leibreich says. ■

Colin Macilwain

A GARDEN FOR ALL CLIMATES

Accustomed to adapting to nature's whims, gardeners are more prepared than most to take on the challenge of climate change. **Emma Marris** asks them what to grow in a greenhouse world.

"This concept that gardening puts you in harmony with nature is a big lie," says Peter Del Tredici, a botanist at Harvard University in Cambridge, Massachusetts. "Gardening is really about preventing nature from doing what it wants to do, which is to destroy your landscape, and gardeners know this at their core. Climate change is just another challenge."

At The Royal Botanic Gardens in Kew, London, the English oaks are ailing. High temperatures and dry conditions over consecutive years have stressed the trees, and wood borer beetles have been taking advantage. "A number of oaks are looking very sad. The weakening of the tree means that beetles come and finish them off," says Nigel Taylor, Kew Gardens' curator of living collections. The leaves, too, are eaten, by the caterpillars of the oak processionary moth, *Thaumetopoea processionea*. To add irritation to injury, these invaders from southern Europe shed hairs that can cause severe allergic reactions in park visitors. "If we get to the stage of a major epidemic, I can see us having to close substantial parts of the park," says Taylor. In ten years, he says, all of London could be affected as the caterpillars become established in northern climates.

Meanwhile, a stroll through Kew Gardens reveals many tender plants and Mediterranean species that would not have been grown outdoors a few decades ago. "The last great winter was 1963," says Taylor. "I remember it from when I was a boy. These days, about a quarter of the plants we grow outdoors would not have survived that winter."

> **"Gardening is really about preventing nature from doing what it wants to do."**
> — Peter Del Tredici

In many places a warmer and less predictable climate seems to be remaking the context in which gardeners sow and reap. Blooming, sprouting and frost times are shifting unexpectedly. Traditional plants are suffering, whereas exotic species are thriving, and unfamiliar pests and weeds are showing up. Gardeners have no choice but to respond to the challenges — and opportunities — offered by their climate-changed gardens.

You can see their responses in the latest trends in British urban gardens: subtropical and vegetable gardening. Olive trees and even tropical avocados have been seen growing in London. In 2006, sales of vegetable seeds in the United Kingdom overtook sales of flower seeds for the first time since the Second World War, according to the Royal Horticultural Society. One reason could be that, for the eco-conscious gardener, home-grown vegetables avoid the carbon emissions associated with importing produce from overseas.

A. MARTIN

Breaking with tradition

Yet the growth of the new brings with it reasonable fears for the old. Some gardeners worry that much-loved traditional species are being adversely affected by the changing climate. And the timing of seasonal events does seem to be shifting. In the Shangai Botanic Garden, cherries and gardenias reportedly bloomed 15–20 days early this year. Del Tredici notes that the annual lilac festival at Harvard's

Arnold Arboretum has been brought forward a week after a couple of Lilac Sundays (as the festival is known) nearly missed the peak bloom. And some of the famous cherries in Washington DC bloomed in January rather than their usual April this year. And one startled man on the street stopped me, saying, "Do you see this? It's not natural!"

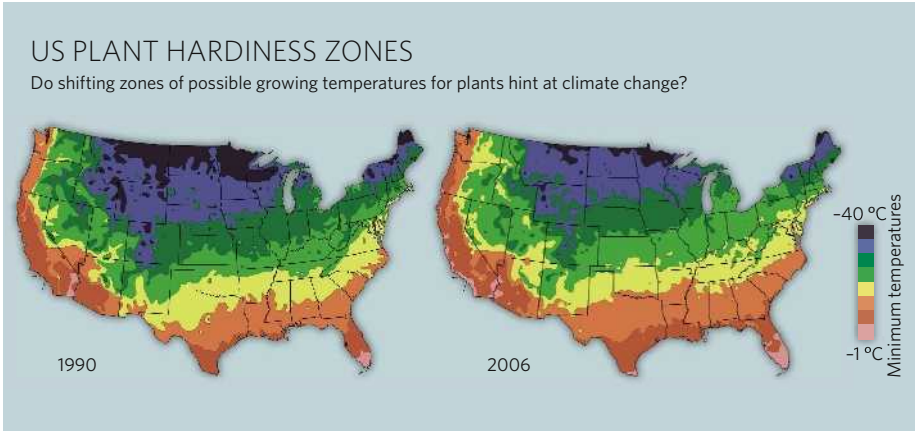
It's not easy to say whether these shifts are caused by global warming or are just the result of natural climate variability. According to Simon Brown, head of climate extremes research at Britain's Met Office Hadley Centre, all that scientists can say for sure about climate change in the United Kingdom is that it increases the probability of extreme events, such as hot, dry summers and mild winters. Events such as the 2003 summer heat wave in Europe, he says, are now at least three times more likely. He adds that most climate models also predict fewer frost days across the country.

Redrawing the map

Changes in the plant hardiness zones used by gardeners to choose what plants to grow are seen by some as evidence of systematic change (see 'US Plant Hardiness Zones'). These zones map the possible growing areas for different plants as defined by regional average annual minimum temperatures. In the United States, the best known map is produced by the Department of Agriculture (USDA). The current map, released in 1990, is based on 15 years of temperature data, and an eagerly awaited revision will be based on 30 years. But impatient with the USDA's progress, in 2006, the national Arbor Day Foundation produced its own update of the 1990 map, based on the 15 years prior to 2006. In it, the zones were shifted noticeably northward, and many chalked that up to climate change.

But USDA spokesperson Kim Kaplan insists that data sets of just 15–30 years are not good enough to diagnose climate change. "Call back in 50 years and I'll let you know," she says. And she points out that temperature is not the only important variable, especially in the activity that is the USDA's main focus: farming. "It is not the temperatures that tend to change what farmers grow, but their effect on moisture," she explains. "One or two degrees isn't enough to affect most plants, but where we are seeing shifting patterns of rainfall, that has a major impact on what farmers can grow." Climate change is expected to increase rainfall in some parts of the world and decrease it in others — with potentially disastrous results for agriculture.

The challenge faced by farmers is much more serious than that faced by gardeners. But precisely because less rides on gardening, horticulture can be seen as a way to experiment with strategies to adapt to changes in the climate, some of which might then have broader relevance.



NATIONAL ARBOR DAY FOUNDATION

"Home gardeners tend to be kind of adventurous," says Peter Raven, director of the Missouri Botanical Garden in St Louis, "so they will continuously be pointing the way to what can be grown." They have the luxury, not available to other land managers, from farmers to city planners, of changing what they do each season. "It is a lot easier in gardening than it is in many other spheres," Raven says, "You can adapt with new plants every year."

And some are already adapting to a warmer and less-predictable climate. "The professional gardening community is beginning to think carefully about what it is going to plant," says Taylor. "Some have even written contingency plans."

Used to suffering from the vagaries of the weather, gardeners might be better prepared than many for the changes that will occur as humanity fills its atmosphere chock-full of heat-trapping carbon dioxide and other gases. "As any gardener knows, the weather was engineered to make us miserable," says Todd Forest, vice-president for horticulture and living collections at the New York Botanical Garden, which recently held a symposium on gardening and climate change. But he adds, "Gardeners love to experiment. They love to try new things. You might be able to grow things in New York that you couldn't grow before. We will look at those opportunities."

FORESTRY COMMISSION

Caterpillars of southern European moths are eating the leaves of English oaks.

Not everyone is so positive. Scott Aker, a horticulturist at the US National Arboretum in Washington DC says that climate change is likely to be mostly bad for gardeners. "I don't believe that global warming is going to allow us to grow things that were previously not hardy enough here," he says. In fact, Aker, explains, because plants go into a state of dormancy for the winter, which is triggered by gradually lowering temperatures, a warm winter and then a cold snap will be much more damaging than the same cold snap after the rigours of a cold autumn and early winter. "We can dispense with the idea that we are going to be growing coconut palms in Washington any time soon," he says.

And some of the knock-on effects of climate change will be too complex to predict. At Kew, they have been watering more in the rash of dry summers they've seen, and all that London tap water is turning their soil alkaline. Changes in precipitation and microclimate will vary, and perhaps the only firm certainty is that the weather will be less predictable.

According to Aker, the uncertain weather wrought by climate change may narrow rather than broaden the range of plants that can be grown in any one place. At least until plant breeders are able to produce tougher varieties. "That is going to be the focus of breeding," says Aker. "The plants that we put in our gardens 20 years from now are going to have to be able to withstand a lot more extremes of

The green gardener

Gardening seems like the ultimate green activity, but it, too, can contribute to greenhouse-gas emissions. Gardens can be incredibly energy- and input-intensive. Trees and other plants capture carbon, but watering, mowing, leaf blowing and using fossil-fuel-derived fertilizers can easily offset this.

"If you are using synthetic nitrogen fertilizers, just purchasing the bag is making a contribution to greenhouse-gas emissions," says David Wolfe, a plant ecologist at Cornell University, in Ithaca, New York. And using too much fertilizer that has a high nitrogen content will release trace amounts of nitrogen oxides. Even tilling may be problematic. "Over-tillage is a big problem — a bad habit that farmers and gardeners get into," says Wolfe. "It over-oxygenates the soil and releases a lot of carbon."

Douglas Kent, an environmental horticulturist and landscape designer in California, offers carbon-neutral and even carbon-negative garden designs. He suggests using ground covers that require low or no inputs of water and fertilizer. Creeping red fescue and sedges are among his favourite lawn alternatives.

Some of his ideas are counter-intuitive. For example, he says that most gardens produce more waste than they can use as compost. Excess wood can be used for borders, and entombing other waste in a landfill makes more sense than letting it decompose, thereby releasing its carbon. And he suggests that, in theory, a lawn that requires heavy inputs might be better for the environment if it were simply paved over. That way you can avoid all the emissions associated with its upkeep.

Climate change may see the end of the 'native garden', a popular trend with many gardeners. As microclimates change, so too will the plants that can survive with few inputs. So gardeners who prefer native plants because they are more environmentally friendly might have to think again. "One of the great myths of gardening is that a native plant is always best adapted to your site," says Todd Forest, vice-president for horticulture and living collections at the New York Botanical Garden.

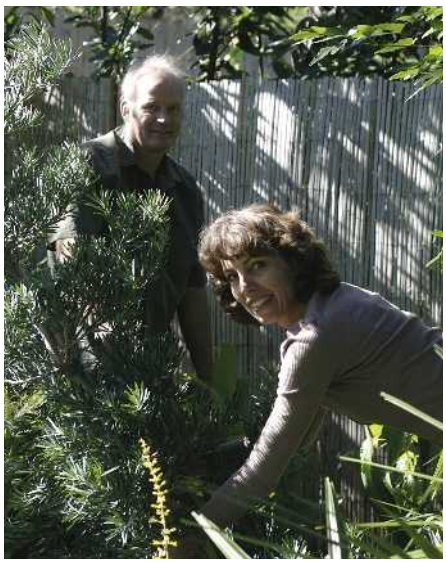
But Kent says that gardening for low emissions, like gardening for unpredictable weather, encourages a beautiful new aesthetic experience. "There is more whimsy, more nature. It is not that apollonian concept of real heavy structure. You kind of usher nature back a little more into your garden." Butterflies and birds prefer these kinds of gardens, he says. "It is exceptionally satisfying, and there is, frankly, less work." **E.M.**

temperature and drought." Apart from changing temperatures and moisture patterns, climate change also expands the ranges for many pests and pathogens. "I would say that perhaps the most significant things affecting horticulture are the new pests and diseases," says Taylor. At Kew they are seeing one or two new pests every year.

Scott Ogden and Lauren Springer Ogden are landscape gardeners who, by virtue of maintaining gardens in two unpredictable climates — Colorado and Texas — are now advising gardeners in heretofore meek and mild climes such as the northeast and northwest. Their advice for handling climate change? Plant more species, so even if some fail others will flourish. They say that if gardeners try to hold on to species they've always grown, they may have to water, fertilize, and generally manage them more ."It's a much more mixed bed," says Ogden. "Instead of maintaining the plants artificially, find the plants that are going to work."

Other advice: forget about relying on long springs to bring out your show-stopping flowering trees — they might bloom in February and then get zapped by a cold snap in March. And watch those formal gardens that rely on broadleaf evergreen hedges. They don't like erratic freezes. Instead, go layered and diverse. Then, "when things ebb and flow there is always something looking good," says Springer Odgen.

The message is to take control by not being too controlling; to worry less about traditional species and to embrace well-adapted species whatever their source (as long as they aren't destructively invasive). The new look in the climate-adapted garden is rambunctious, diverse and more like wild spaces.



M. NOWACKI

Scott Ogden and Lauren Springer Ogden run a consultancy for green gardening in Texas.

Up to date

No matter what their local climate does, gardeners will notice. "If you ask a gardener what the ten-day forecast is or whether it has been a wet or dry fall, chances are that they will know," says Forest. In a sense, every year is a new mini-experiment in each garden. And in some regions gardeners are being asked to put their famous attention to detail to scientific use by recording and reporting data. Phenologists, who study seasonal phenomena, are enlisting citizens to record data in projects such as the US-based Project Budburst, the United Kingdom's Nature's Calendar, the Netherlands' De Natuurkalender and Canada's NatureWatch.

Kayri Havens, a conservation biologist at the Chicago Botanic Garden in Illinois, helps run Project Budburst. She says that in its pilot year, about two-thirds of their 1,800 observations of blooming times were from children under 12. Havens hopes that the data will be used to predict where plants may need to migrate so that their blooming coincides with their pollinators' cycles. Data from an older volunteer project run by the University of Wisconsin, Milwaukee, which monitors lilac bloom times, have already been used in more than two dozen scientific papers.

Gardeners are moving into a space where many others are still loath to go. When it comes to climate change, says Raven, gardeners can afford to experiment. They accept climate change as fact, and they work with it (see 'The green gardener'), some even do so cheerfully.

"I call this a brave new ecology," says Del Tredici, "The reality is that climate change is already happening, so we have to learn how to live with it." This approach might be a model for managing other activities, beyond the backyard. After all, says Stephen Hopper, head of Kew, "Some people argue that the world is managed so much that we are all gardeners." ■

Emma Marris is keeping an eye on when the redbuds bloom in Columbia, Missouri.

Pieces of the puzzle

After decades of war, looting and destruction, Afghanistan's archaeologists are scrambling to restore their country's cultural heritage. **Rex Dalton** visited Kabul to see how they are faring.

From a hillside overlooking Kabul, a dozen Afghan archaeology students have a monumental vista of their nation's ancient heritage. Domes of tombs of past kings dot the skyline; a stone wall topped with battlements snakes along a ridge; and nearby looms the fifth-century AD fortress Bala Hissar, site of countless battles and events, including the massacre of a British envoy and his staff about 125 years ago and the retaliatory series of public hangings of Afghans.

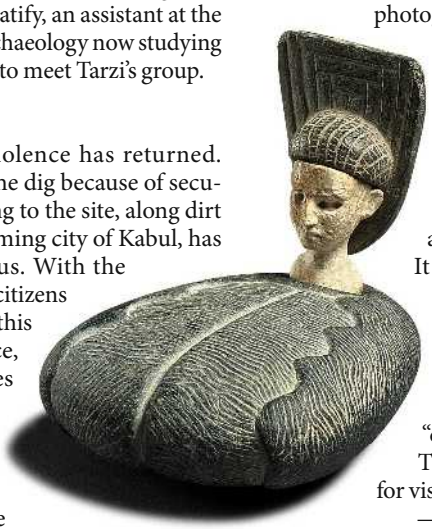
But it is a less dramatic, century-old home site, on a knoll on the hillside, that is getting the attention today. Afghan archaeologist Zemaryalai Tarzi, of Strasbourg University in France, has brought the students here to teach them basic excavation skills. The group was to have been the inaugural class of Afghanistan's first graduate programme in archaeology at Kabul University. But instead it has turned into an impromptu field school; the challenges of setting up a master's programme in the country are too great now.

It is six years since the fall of the Taliban regime, the force that routinely and ruthlessly smashed artefacts it deemed idols. And although normality is returning to many parts of Afghan life, archaeologists are still struggling to recover the country's heritage and rebuild its academic community. The threat of violence keeps many researchers from doing field projects, as funding agencies often ban archaeologists from going out to sites. Looted artefacts are being recovered, but only slowly. Many artefacts that did make it through the strife — including a priceless collection of gold relics, called the Bactrian hoard — are now being exhibited abroad, although critics claim that Afghanistan is not being sufficiently compensated.

Taliban memories haunt nearly every aspect of Tarzi's instructional dig. The excavation site lies below a steep-walled, rocky gorge that cuts to the top of a high ridge and also holds a spring, a source of fresh water for the poor community just down the hillside. There, supplied with water that could be traded for food, the Taliban set up a command post from which to strategically control the western approach to Kabul. "No one could come here during Taliban days," says Hafiz Latify, an assistant at the Afghan Institute of Archaeology now studying in Greece, as we climb to meet Tarzi's group.

Kidnap threat

But the spectre of violence has returned. Tarzi has suspended the dig because of security concerns; travelling to the site, along dirt alleys through the teeming city of Kabul, has become too dangerous. With the kidnapping of French citizens in Afghanistan earlier this year, officials in France, where Tarzi now lives and works, say he can conduct studies next summer only in Bamiyan, where the situation is more secure. Over the years, Tarzi's excavations at Bamiyan, about 125 kilometres west of Kabul, have yielded an array of artefacts, including life-like sculpted heads modelled after individuals from the past two millennia. But he wants to return to Kabul so that he can resume his teaching efforts.



BOISGIRARD

One of two Bactrian sculptures sold at auction in France earlier this year.

In some areas of Kabul, though, reconstruction is already under way. West over the ridge from Tarzi's instructional dig, the Afghan national museum in Kabul has undergone a transformation. The museum was ransacked under Taliban rule, and statues were pounded into smithereens in a rampage that did not garner as many headlines as the destruction of the huge Buddha statues at Bamiyan. By the time they left, "the museum was a depressing ruin — no roof, no glass, everything broken into little pieces", says Gitta van Buuren, a Dutch cultural anthropologist who has visited Kabul since 2003 to document the city's recovery photographically.

Today, the national museum is completely refurbished, the beneficiary of aid from UNESCO (the United Nations Educational, Scientific and Cultural Organization) and numerous countries. It sits in stark contrast to the Darul-Aman Palace, a bombed-out stone edifice still majestic on a nearby hilltop. The museum restorers, says van Buuren, "did an amazing job".

The museum is now open for visitors — although few come — and in a laboratory on the second floor, museum staff spread out broken pieces, putting artefacts and statues back together like jigsaw puzzles. The pieces fill buckets. But the staff is turning them back into life-like forms, and the restored statues are making their way downstairs along with other salvaged objects. Omara Khan Masoudi, the museum's director, says the restoration team is making progress, but is short-staffed when it comes to skilled workers. "We need more Afghans — or any scientist — to help," he says.

Facing such realities, Afghanistan has turned to touring some of its most precious artefacts in international museums. Afghanistan's key treasures include the Bactrian hoard — about 21,600 gold coins, ornaments, pieces of jewellery and funerary relics — discovered in northern Afghanistan in 1978 at a 2,000-year-old burial site called Tillia Tepe. They reflect a mélange of styles — from Greek to those of Asian tribes.

When the Taliban tightened authoritarian screws, archaeologists worldwide feared these relics were lost. But in 2003, the Bactrian hoard

R. DALTON

Students practise techniques for archaeological excavation on a hilltop overlooking Kabul.

T. OLLIVIER/KABUL MUSEUM

On the road: more than 200 spectacular Afghan treasures are currently touring the world.

emerged from a vault of the central bank in Kabul, where it had been successfully secreted away from the Taliban. After inventory and cataloguing, the gold artefacts went on the global museum circuit.

In the past year, more than 200 artefacts from the Bactrian collection have been exhibited in Paris, and Turin in Italy. On 22 December, the exhibition is to open at the Nieuwe Kerk in Amsterdam, the Netherlands, in what is being called a 'blockbuster show', before moving to a US tour that takes in Washington DC, San Francisco, Houston and New York. But simmering behind the glitz is anger and resentment about whether the Afghans are being properly compensated. "The Afghans were taken advantage of," charges Lynne Munson, a former deputy director of the US National Endowment for the Humanities, who helped arrange endowment funding for the Bactrian inventory.

The European exhibitions typically paid about €150,000 (US$220,000) to Afghanistan; in Paris and Turin, about 130,000 people visited the exhibition, at an admission fee of €8 apiece. The four upcoming US exhibitions — potentially the richest revenue producer — are to provide the Afghans with a total loan fee of $1 million. The Afghans will also receive 40% of merchandise sales, after expenses, for the US tour.

Making the deal

Munson argues that the Afghans should have received substantially more, and she worries that they will see nothing from the merchandise deal given the way it is structured for payments after expenses. She blames the US tour's organizer, the venerable National Geographic Society, which will receive any additional monies from the tour. "They took advantage of the Afghans for their own selfishness," she says.

But Terry Garcia, an executive vice-president of the National Geographic Society, says his organization worked hard to make sure the Afghans were getting a good deal, and modelled US financial arrangements on those of the European tour. "Throughout every step of the process, we have responded to the needs and wishes of the Afghans," he says.

Munson also says that she recently learned that the document lending the artefacts was signed on behalf of the Afghan government by Omar Sultan, who took on the responsibilities of acting minister for the Afghan culture ministry after the minister was injured in a bombing. But Sultan has also worked as a paid consultant to National Geographic. "It's a conflict of interest," says Munson. The arrangement has been debated and criticized in the Afghan parliament, although Garcia says the negotiations had the full support of the Afghan government.

The need for funds is desperate, says Ana Rosa Rodriguez, executive director of the Society for the Preservation of Afghanistan's Cultural Heritage. Her group, based in Kabul, trains staff and raises funds to protect sites by actions such as trying to restrict detrimental development at the 'city castle' of Bala Hissar. More money would translate into more conservation, she says.

But some Afghan supporters say that sending the nation's artefacts on tour is a good idea — as there is limited security to protect them at the Kabul museum. Afghan artefacts of suspicious provenance — potentially sneaked out during the Soviet or Taliban years, or even since — regularly show up for sale at international auction houses. In April, the International Council of Museums (ICOM) in Paris presented a 'red list' guide to Afghan artefacts that may appear for sale illegally. Afghan treasures that have been up for auction have been scrutinized, but the list hasn't yet resulted in any seizures by law-enforcement agencies such as Interpol, says ICOM's Jennifer Thevenot. This isn't surprising, as both the ICOM and Interpol have skeleton staffs for looking into questionable artefacts. If an auction house has legitimate-looking documents, an artefact is likely to be sold with little inquiry.

In November, for instance, the Boisgirard auction house in Paris offered two Bactrian sculptures at prices of up to US$100,000. Thevenot says that questions were asked, but no action was taken because the house had documents. No one contacted Afghan experts such as Tarzi, who helped to draw up the red list.

The items were sold, disappearing into a private collection much as they might once have gone down the old Silk Road. Those trying to save Afghanistan's culture hope that it doesn't all follow that route. ■

Rex Dalton is a US West Coast correspondent for *Nature*.

D. PARKINS

SCANNING PSYCHOPATHS

Are their brains not wired to feel what others feel, or do they just not care? **Alison Abbott** joins researchers looking into normal neurobiology through the scope of psychopathy.

It is a rare event that patient 13 is let out of the high security Dr S. van Mesdag Clinic in Groningen, the Netherlands, and he is making the most of the attention he is getting. Already, the prison guards have had to accompany him from the University of Groningen's functional magnetic resonance imaging (fMRI) scanner to the toilet four times in two hours. The guards indulge him with a shrug. Research psychologist Harma Meffert, who has recruited him for her study, is just as tolerant. That can't be easy, given that she has to spend at least 20 minutes resettling him into the scanner after each interruption

Wearing nothing but blue cotton surgical pyjamas and a constant smile, patient 13 doesn't seem to present much of a threat. In fact with his jewellery removed and his tattoos covered he looks decidedly small and vulnerable. But no one is forgetting why he was recruited to Meffert's study. Patient 13 has scored the maximum possible on the Psychopathy Checklist-Revised (PCL-R) rating scale, the ubiquitous tool psychiatrists use to identify the personality and behavioural traits that define the clinical syndrome 'psychopathy'. Lack of empathy is a key feature.

"Do they do terrible things to other people because, unlike most of us, they do not share the pain they inflict?"
— Christian Keysers

As it happens, Meffert's lab chief, Christian Keysers, the 34-year-old director of the university's neuroimaging centre, is not primarily interested in psychopathy per se. The major focus of his research is empathy — the way we can't help feeling awful when we see a loved one cry, or can't stop our stomach sinking when someone's face darkens in anger. One major theory holds that we understand what another person is feeling by activating the same neural circuitry in our brains that activates when we are experiencing that emotion first hand.

To investigate this trait, Keysers is comparing 'normally' empathic people with those who lack empathy, such as people with autism, and psychopaths. He suspects that psychopaths may be able to recognize emotions in others but that they are also able to disconnect that recognition from their own emotions. "Our question is: do they do terrible things to other people because, unlike most of us, they do not share the pain they inflict?" says Keysers. His sophisticated trial design is intended to test whether this is the case (see 'Letting fingers do the talking').

Although not all diagnosable psychopaths are criminally inclined or in prison, places such as the Groningen clinic serve as a concentrated source. And they provide screening. The PCL-R scale is practically the only tool available for this purpose. In PCL-R assessment, specially trained psychiatrists discuss hundreds of issues with the patient during semi-structured interviews. On the basis of these interviews, and information about past behaviour from independent sources, such as social workers' reports, they build up a four-part assessment. The headings are: 'interpersonal', covering behaviour such as manipulativeness and lying; 'affective', covering irresponsibility and lack of empathy and remorse; 'lifestyle', tracking impulsivity and need for stimulation; and 'antisocial', which looks for records of things such as juvenile delinquency. Those on the receiving end of the assessment find it tiresome.

It is, of course, not easy to put together a group of imprisoned psychopaths for an academic research project, but the Dutch Ministry of Justice provides generous access. "We have a legal duty to try to treat all those criminals who

are found guilty but not responsible for their actions due to insanity," says Jacqueline Hochstenbach, a department head at the ministry.

Although the Groningen project doesn't aim to treat or cure psychopathy, there's a general sense even among the subjects that such basic research could at least help illuminate what is wrong. For psychopaths who are deemed dangerous, there are no therapeutic options. "We know there is no effective treatment for psychopathy," says Hochstenbach. Pharmaceuticals don't help and those who receive behavioural therapy have a higher — not lower — rate of recidivism.

In 2004, like other countries, the Netherlands institutionalized PCL-R testing in forensic psychiatric centres, where it is used as a risk-assessment tool for patients being considered for parole. Developed over the past three decades by psychologist Robert Hare from the University of British Colombia in Vancouver, Canada[1], it has proved to be a powerful predictor of the likelihood that a criminal will reoffend.

NEUROIMAGING CENTER, UNIVERSITY MEDICAL CENTER GRONINGEN

Christian Keysers (right) and Harma Meffert want to dissect empathy by scanning psychopaths.

A qualifying score

To qualify for the empathy study, participants must score higher than 30 on the PCL-R scale, out of a maximum possible score of 40. Qualifiers are told they will participate, but not when — to give no opportunity to plan escape. On the morning of the test, they are asked to confirm their consent and Meffert goes through the protocol again in more detail. She does not, however, explain the detailed scientific aims of the study in case the subjects try to manipulate the outcome.

Inside the clinic, Meffert is often alone with her subjects but wears an alarm around her neck. "Once I pressed it by accident and was amazed to find myself surrounded by several guards who seemed to spring from nowhere within seconds," she says. "I feel safe." Meffert is a calm person, who works well with her subjects by talking and listening to them seriously. But she says that psychopathic people can be very tiring to work with because they command, and need, intense attention.

On the morning of his test-day interview, patient 13, although taken by surprise, is looking what must be close to his best. His hair and beard are fashionably trimmed, and his clothes are casual but coordinated. Walking to the small interview room, he says he wishes he had more notice, but he is laughing. He listens to Meffert's detailed explanation of how the day will run and gives his agreement.

"Some psychopathic features are not necessarily a bad thing for society — in some professions they may even help."
— Robert Hare

An hour later he is on the road, in an armoured van. No metal is allowed near the scanner. Even his tattoos nearly ruled him out as a subject, but they are small, and also recent enough that the red in them is likely to be from newer, iron-free dyes that won't affect the imaging. The guards don't carry guns, the rod fitted down patient 13's trouser leg, preventing him from running, is security enough. Handcuffs are made of a special non-metal material.

Patient 13 doesn't seem to think much of the experiment itself. The sequence of film clips, which are projected inside the scanner directly above his face, only run for ten minutes or so. But he finds it hard to concentrate and his eyelids, observed remotely by the researchers in the adjacent control room, begin to droop. Meffert runs the clip again — the experiment requires the subject's full attention.

Most of the others who have taken part in the study were much more compliant and easier to handle in the scanner than patient 13 — often they are more cooperative than the average student volunteer, says Meffert. All the subjects seemed to find the experiment to be nonsense. "It was stupid, boring," says inmate Willem Boerema (not his real name), who claims to have taken part only because he likes Meffert. Then, contradicting himself, he adds that "if they say the study can help people then it's good".

Boerema, smart, articulate and multilingual, has a PCL-R rating of 35 — and a big problem with the term 'psychopath'. He views it as a fashionable label abused by the judicial system to keep people like himself from being released. "The courts look at your PCL-R rating and add two years to your sentence, then another two years, and then another."

Damaging label

When he entered the prison five years ago, Boerema says, 'borderline personality' was the fashionable term, and his designated pigeonhole. "The psychopathy label is more damaging though — it prompts everyone to see you as a potential serial killer, which I could never be." (Note, in reporting this article it was agreed that inmates' crimes would be neither asked about nor reported on.) But Boerema also wears the score as a badge of honour: "I think my high psychopath score is a talent, not a sickness — I can make good strong decisions, and it's good to have some distance with people."

There is some truth in this, says Hare. As well as developing the PCL-R, he has also developed a shorter version suitable for screening the general population (PCL-SV, with the SV standing for screening version). He has used it to estimate that maybe 1% are psychopathic, even if they have never committed a crime, according to research presented recently at a meeting on psychopathy research. "Some psychopathic features are not necessarily a bad thing for society — in some professions they may even help," says Hare. "Too much empathy, for example, on the part of a police officer or a politician would interfere with the job."

In theory, scientists like Keysers could recruit high-scoring psychopaths from the general population as control subjects for studies on empathy. But identifying enough of them would be extremely time consuming. Some scientists without access to the captive population in prisons — a fifth of whom may be psychopathic according to Hare[2] — have turned to populations on the outside with specific behavioural problems.

Neuroscientist Jorge Moll, for example, from the Labs-D'Or Hospitals network in Rio de Janeiro, Brazil, screened and recruited 'troublesome' outpatients of a civil psychiatric centre for his ongoing neuroimaging study to identify the neural circuits involved in moral judgement. Drug use, which can interfere with results, is a bigger hazard in those outside prisons than those inside, he concedes, "but the standards of security in Brazil don't make a prison study feasible here".

James Blair at the National Institutes of Health in Bethesda, Maryland, gets around the drug problem by using children with behavioural problems who score highly on the PCL-SV rating, and whose parents have responded to his advertisement. Kent Kiehl, now at the University of New Mexico in Albuquerque, has worked with parole populations in Connecticut, but found the subjects so unreliable that he spent three-quarters of his time getting them to keep appointments. New Mexico is one of several US states that, like the Netherlands, are keen to promote psychopathy research. Kiehl has made the most of the supportive environment and developed a mobile fMRI machine to conduct a dozen or so different studies — from empathy and moral reasoning to cognitive function — on 300 inmates. "Going into the prison means you can get many more subjects than would be possible by bringing them out individually with all the arrangements that requires," he says. "Larger subject numbers mean a more definitive study."

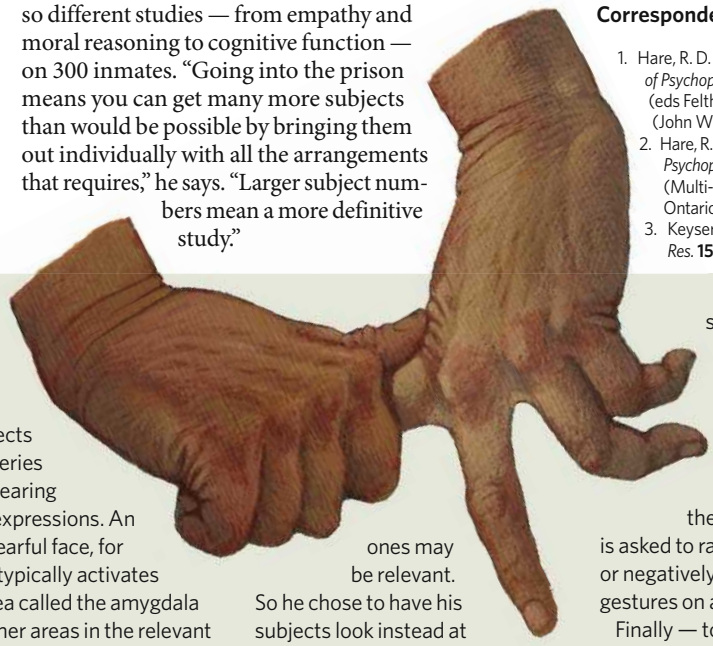
"I think my high psychopath score is a talent, not a sickness."
— Willem Boerema

All of these psychopathy researchers believe that their work will lead to a level of understanding of the condition that could eventually lead to a treatment. Keysers does too — even though his prime motivation in recruiting psychopaths was to support his empathy research. He now finds himself "fascinated by the phenomenon of the untreatable psychopath", and also convinced that there will one day be a fix. Patient 13, meanwhile, has finished his test day wearing the same smile he set out with. If a form of therapy were ever to emerge, it is not clear whether people like him — who do not consider themselves sick — would be willing to take it. ■

Alison Abbott is *Nature*'s Senior European Correspondent.

1. Hare, R. D. in *The International Handbook of Psychopathic Disorders and the Law* (eds Felthous, A. & Sass, H.) 41–67 (John Wiley, Chichester, UK, 2007).
2. Hare, R. D. *Manual for the Revised Psychopathy Checklist* 2nd edn (Multi-Health Systems, Toronto, Ontario, Canada, 2003).
3. Keysers, C. & Gazzola, V. *Prog. Brain Res.* **156,** 379–401 (2006).

D. PARKINS



Letting fingers do the talking

'Normal' people run simulations in their brain mirroring what others feel. This means that when they see someone expressing (through facial expressions and gestures) a particular emotion, some of the same circuitry is activated in their own brains that would be activated if they were feeling that emotion themselves[3]. This circuitry feeds into cognitive and emotional processes. So the empathy is both an intellectual and an emotional process.

Psychopaths have a number of defining features, one being a pronounced lack of empathy. Do these individuals fail to mirror the emotions of others, or do they detach the emotional component when the mirroring process happens? Christian Keysers, the director of the Neuroimaging Centre at the University of Groningen in the Netherlands has designed an experiment to address this question.

Most researchers studying empathy or specific emotions have their subjects look at a series of faces wearing different expressions. An angry or fearful face, for example, typically activates a brain area called the amygdala among other areas in the relevant neurocircuit. This activation can be seen indirectly using functional magentic resonance imaging, which measures changes in the brain's regional blood oxygenation.

Keysers saw limitations to this paradigm for his own research. "With faces you are optimizing the brain responses for particular areas — and those areas are going to be the only ones where you'll be able to see any potential difference between psychopaths and non-psychopaths of the same age and educational background," he says. He wanted to develop a paradigm where there would be activation in many brain areas, so that differences in activation could be seen in circuits without preconceived theories about which ones may be relevant. So he chose to have his subjects look instead at hands interacting in ways that convey different emotional messages.

With his team he made videos of two hands interacting for just a few seconds. One hand strokes the other, for example — a positive, pleasurable gesture. Or it twists the finger of the other to cause pain. Or it simply shakes the hand, an emotionally neutral gesture. In a more sophisticated example, one hand gently approaches the other only to be slapped aggressively away.

They made a ten-minute sequence of the recorded gestures, with eight-second gaps between each. The experimenter shows the sequence to the subject as he lies passively in the scanner. Then the subject has to watch the movies again, but this time must put himself in the shoes of either the victim or the perpetrator. Then, once outside the scanner, the subject is asked to rate how positively or negatively he perceived the gestures on a scale of one to five.

Finally — to identify brain circuits involved when the subject is directly exposed to emotionally laden hand gestures, rather than just watching them — the experimenter repeats the same hand interactions on the hands of the subject — stroking, twisting, shaking, slapping away each of their hands in turn while inside the scanner. The subject again responds using the sliding scale.

"The paradigm with the interacting hands activates many parts of the brain, including those involved in processing movement, touch and emotion — so we hope we'll be able to identify areas that are activated by the normally empathic, but not activated by those who lack empathy," says Keysers. **A.A.**

All fishing nations must unite to cut subsidies

SIR — The threat of overfishing to world fisheries is well documented, but not enough attention has been paid to government subsidies as an important factor in their decline. Subsidies, or government payments to the fishing sector, estimated at US$30–34 billion a year, are key drivers of the unsustainable exploitation of the world's depleted fish populations. Fish are the main source of protein for one fifth of the world's population, but global fishing fleets are more than double the size the oceans can support.

If fisheries are to become sustainable, overfishing subsidies must be significantly reduced (U. R. Sumaila *et al. Fish. Res.* **88,** 1–4; 2007). Unfortunately, unilateral action by individual countries may not work, because their fisheries could then be at a disadvantage in the competitive global market for fish; also, fish do not respect national boundaries and fishing fleets operate worldwide. The only effective approach to the subsidy problem is through multilateral action, in which all fishing nations end or reduce these subsidies under similar rules.

The World Trade Organization (WTO) has 151 member countries and a mandate to level the trade playing-field for every country. It is in a unique position to tackle the global problem of overfishing subsidies and to move fisheries towards sustainability, because it is the only global institution, apart from the Convention on International Trade in Endangered Species of Wild Fauna and Flora, that has mechanisms in place to enforce its agreements.

The WTO is at present drawing up terms, including terms on how to police fisheries subsidies, in the Doha round of negotiations. We urge the WTO to seize this opportunity now, to forestall the predicted collapse of the world's wild fish populations.

U. Rashid Sumaila*, Daniel Pauly†
*Fisheries Economics Research Unit, Fisheries Centre, The University of British Columbia, †Fisheries Centre, The University of British Columbia, AERL Building, 2202 Main Mall, Vancouver, British Columbia V6T 1Z4, Canada

Debate over flood-proofing effects of planting forests

SIR — William Laurance, in his News & Views article 'Forests and floods' (*Nature* **449,** 409–410; 2007), highlights a paper by C. Bradshaw and colleagues, claiming that it provides correlative evidence that native forests reduce the frequency and severity of floods in developing countries.

The 'forest and floods' debate goes back at least to the nineteenth century. Now forest hydrologists generally agree that, although forests mitigate floods at the local scale and for small to medium-sized flood events, there is no evidence of significant benefit at larger scales and for larger events.

Laurance also recognizes the omission of extreme events in the Bradshaw analysis. But we argue that this seriously weakens the policy importance of the results. It is these extreme events that matter: economic damage and loss of life grow exponentially with flood magnitude. The authors also excluded Chinese data because of outliers — an unfortunate omission, given that China has undergone large changes in forest cover where the flood 'signal' should be strong.

Laurance did not discuss Bradshaw and colleagues' other conclusion, namely that increasing the number of forest plantations can lead to longer and more frequent floods. For China, where initiatives such as the sloping-lands conversion programme — heavily promoted on the basis of flood-mitigation benefits — are leading to forest plantation over areas comparable to those of afforestation by the rest of the world put together, the conclusion is particularly important. Other studies have warned that forest-management activities can aggravate flood risk (see J. A. Jones and G. E. Grant *Water Resources Res.* **32,** 959–974; 1996).

Ian R. Calder*, James Smyle†, Bruce Aylward‡
*Centre for Land Use and Water Resources Research, Newcastle University, Newcastle NE1 7RU, UK
† 149 East Rosewood Avenue, San Antonio, Texas, USA
‡Ecosystem Economics LLC, PO Box 2062, Bend, Oregon 97709, USA

Motivation needed to cure lifestyle diseases

SIR — David A. King and Sandy M. Thomas in their essay 'Big lessons for a healthy future' (*Nature* **449,** 791; 2007) describe how, in Western societies, "growing recognition of how science can contribute to health, well-being and the economy" is leading to governmental attempts to control public health. Their Foresight study suggests that by 2050, in the United Kingdom, about 60% of men, 50% of women and 25% of children will be obese, and that the associated chronic health problems will cost an additional £45.5 billion (US$93 billion) a year. To combat this, the authors propose a strategy similar to that used for climate-change policy, in which six chief advisers (and their entourages) of government departments developed a 'top-down' strategy.

Such a paternalistic approach is reminiscent of past attempts to eradicate cholera and smallpox by better sanitation, housing and vaccination. However, today's lifestyle-related diseases, such as obesity, type-2 diabetes mellitus and cardiovascular disease, are different. They develop slowly, in response to chronic food intoxication and lack of exercise, causing a long-lasting imbalance between excess energy intake and insufficient energy expenditure. And they require a different approach.

Paternalistic health-care systems will be of little avail because they fail to activate the individual's motivation to care continuously for his or her own health, which is necessary for success. To secure such motivation, there would have to be annual financial incentives for those complying, with set targets as to body weight and physical strength. Such incentives could be tax breaks, or a partial annual cash refund of private health-insurance premiums. The aim would be absence of lifestyle-related disease at the age of 60. An addressable target group for such incentives could be people aged 35–55 who care for children and/or elderly family members, as additional pressure from their dependants might help to secure compliance with health targets.

Werner Waldhäusl
Department of Internal Medicine III, Medical University of Vienna, Währinger-Gürtel 18–20, 1090 Vienna, Austria

Educational success must start in Pakistan's schools

SIR — Your Editorial 'The paradox of Pakistan' (*Nature* **450,** 585; 2007) rightly credited General Pervez Musharraf's regime for increased spending on science and research, which may be threatened by the political decline of Musharraf. I would like to add that the fall of such an ambitious scientific enterprise is built into the educational system of the country at primary and secondary levels. Universities simply cannot produce quality research and education if fiscal resources are not accompanied by quality human resources.

The students at Pakistan's universities come from a vastly neglected school and college system. Unless the lower-tier educational institutes are able to produce quality pupils, the money spent on universities is going to be wasted. The school education system must be thoroughly overhauled so that universities can make the best use of available resources.

Masroor Bangesh
Institut für Anorganische und Analytische Chemie, Friedrich-Schiller Universität Jena, Carl-Zeiss Promenade 10, 07743 Jena, Germany

Contributions to Correspondence may be submitted to correspondence@nature.com. We also welcome comments and debate at http://blogs.nature.com/nautilus.

BOOKS & ARTS

Small matters, big issues

Science books for children are thriving, partly because of the competition from new media.

Harriet Coles

The printed page was the primary source of information and inspiration in our childhoods. This is not so for today's children. Electronic media now compete with science books; the Internet and CD-ROMs being particularly well suited to the presentation of non-linear, non-narrative information. In response, authors and publishers have had to explore new formats for science books for the young. This special issue examines some of the exciting results.

The best science books engage, educate and inspire a child. And the most ambitious of these discuss the concepts and process of science, not just the facts. The educational brief is complicated by the ways in which different children respond to new information. For children who will sit for hours reading encyclopaedias, a new generation of reference books, rather than following the alphabetic imperative, present facts in categories and familiar contexts (see 'The scene is set', page 952). For children who are unlikely to seek out a book about science, some authors use stories (see 'Hawking's fact and fiction' and 'Stones, bones and stories', page 949, and 'Star Tales', page 950) and others use humour and questioning to make science more familiar (see 'Mathematics not Shopping', page 951).

Books that come with Internet links or CD-ROMs (see 'To bodily go ...', page 951) usually fail to integerate the printed and electronic page. Print endures much longer than the average URL, and web links in books even a couple of years old are often obsolete. More successful are the increasingly inventive pop-up books that encourage children to explore scientific knowledge. By using few words to convey each point and hiding information under flaps and in illustrations, pop-up books are both the precursors of and a response to multimedia's selectable links and embedded content.

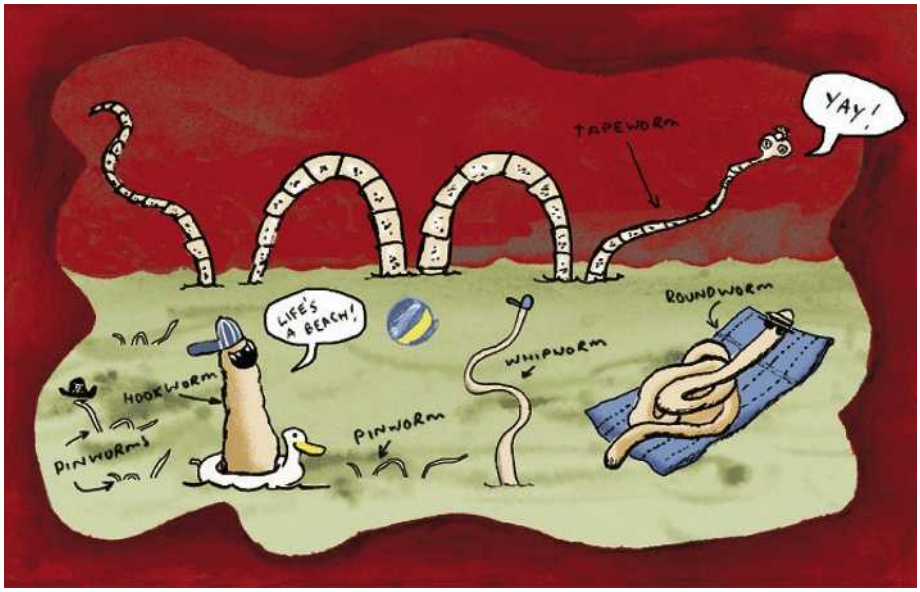
Whatever the format, a truly inspirational book needs an author with a command of and love for the subject. *Why is Snot Green?* (8–14 years) by Glenn Murphy is an excellent example. Set out as a lively dialogue between Murphy and the reader as his questioner, it makes learning effortless and enjoyable. The questions range from space to biology and are so astute and well pitched that you soon forget you didn't ask them. The answers are satisfyingly rooted in everyday life. Most important, Murphy shows that questions are at the heart of the scientific approach. It is a modest book with no gimmicks and minimal graphics. Modest books rarely attract children, however, and so depend on an interested adult to get the child engaged.

Left to choose for themselves, children are drawn to attention-grabbing devices and publishers are becoming expert at exploiting this. We took six books into Brill Church of England Combined School in Buckinghamshire, England (see www.nature.com/nature/podcast). The class of 9–10-year olds was immediately attracted to *The Gooey Chewy Rumble Plop Book* by the grotesquely stretchy, sticky three-dimensional latex tongue set into its front cover.

This imaginatively engineered pop-up book follows the journey that food takes through our bodies. It mixes intelligent, easily digestible text with imaginative pictures. On opening one page, a larger-than-life mouth is projected at the reader; on another, there's an extendable small intestine. The gimmicks, although initially a distraction, are a real treat to explore — revealing lots of engaging facts under flaps.

Do all these bells and whistles help the child's understanding? When Charlotte, aged nine, was asked what she had learned from *The Gooey Chewy Rumble Plop Book* — her favourite — she replied "nothing", and I suspect that her first encounter with the book was dominated by its visual inventiveness. The intention presumably is that a child taken with the physical experience of the book will return to it again and again, learning at each reading.

Over the past ten years, author Nick Arnold has made an industry out of appealing to children's sense of humour with the madcap, gory *Horrible Science* collection — selling 4 million copies in Britain alone. His latest, *The Horrible Science Annual 2008* (8+ years) is crammed with facts — from the cell biology of neurons (the nervous system is described as a telephone exchange) to the fate of Einstein's brain after he died. These are delivered with energy and imagination in cartoons, summaries and activities. April, aged ten, said "they make you want to read on because they say funny things"; Joly, aged nine, described them as: "Perfectly presented, absolutely brilliant, very funny."



N. DAVIES *WHAT'S EATING YOU?* ILLUSTRATED BY N. LAYTON, WALKER BOOKS.

Even more impressive is the collaboration between author Nicola Davies and illustrator Neal Layton. Their infectious enjoyment of zoology always yields fresh perspectives. Layton's illustrations are a natural foil to Davies's direct and surreptitiously informative books. Their most recent work, *What's Eating You?* (8+ years), is a delightfully quirky text on parasites in the language of everyday life (pictured). The reader gains a good sense of habitat, life-cycles, malaria, the importance of grooming and even impalas' bottoms. Davies and Layton get the balance right between wonderful presentation and inspiring wonder at nature. As nine-year-

old Luke put it: "The pictures are great and the parasites are amazing."

The frontier for pioneering authors and publishers, beyond the ever-popular dinosaurs and extant animals, is giving a feel for the nature of scientific research. *Famously Foul Experiments* (8–16 years) succeeds splendidly. Nick Arnold explores key experiments in the history of science using simple activities for the reader to do at home, short biographies of the scientists who first established the principles and a pithy explanation of the concepts. We are given Hubble and expanding balloons, Darwin and a game of natural selection using coloured paperclips, and Ibn Al-Haytham and the pinhole camera, to name just a few.

In *The Global Garden* (6–12 years) by Kate Petty and Jennie Maizels, garden gnomes bearing schematic molecules of carbon dioxide and water appear on a page of pop-up plant nutrition to illustrate photosynthesis in this delightful book on the origins of food and the global economy. It is exciting to see authors effortlessly including plant biochemistry and physiology as part of a broader story, much as they are in life. Six-year-old Nell and 13-year-old Floss were both delighted by the gnomes — illustrating how the very best books appeal to readers of all ages. Similarly, Arthur Kornberg — yes, of DNA synthesis fame — spans the generations in *Germ Stories*. This collection of cautionary verses on microbiology were originally written for his grandchildren and many of the rhyming couplets are a delight to child and adult.

Some cultural commentators say that books are enjoying their final years of supremacy. Whether this is the case or not, recent competition from the new media has only been a good thing for children's science publishing. Books such as the ones reviewed here make the case for a strong future for the printed page. ■

Harriet Coles was formally Arts and Books Editor at *Nature* and is commissioning editor for this children's science book issue.

Why is snot green?
by Glenn Murphy
Macmillan: £4.99, $10.31

The Gooey Chewy Rumble Plop Book
by Steve Alton & Nick Sharratt
Bodley Head: £9.99, $17.99

Famously Foul Experiments
by Nick Arnold
Scholastic: £5.99, $12.38

Horrible Science Annual 2008
by Nick Arnold
Scholastic: £6.99

What's Eating You?
by Nicola Davies & Neal Layton (illus.)
Walker Books: £7.99, $12.99

The Global Garden
by Kate Petty; Jennie Maizels (illus.)
Eden Project/Random House: £12.99/$14.99

Germ Stories
by Authur Kornberg; Adam Alanz (illus.)
University Science Books: £22.50

SPUD GOES GREEN, EGMONT PRESS. ILLUSTRATION BY N. BAINES.

Young planet-savers

Tom Standage, with help from Ella (7½)

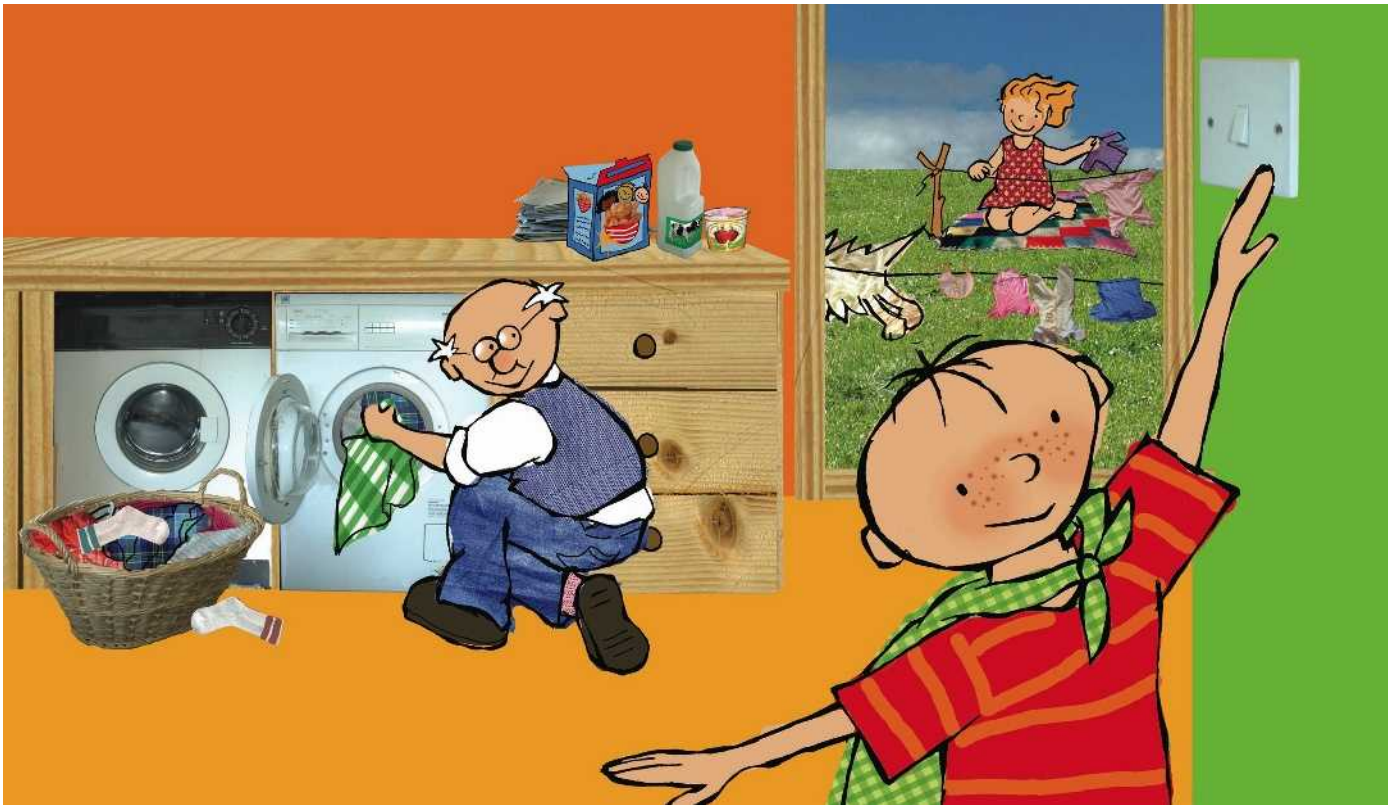
Their parents grew up in the shadow of a possible nuclear war. Children today are growing up in the knowledge that the environment is in peril — and that some actions make things better whereas others make things worse. Last year, my daughter Ella, then aged six, began to ask whether various activities, such as bouncing on her trampoline, "made global warming" or not. If Ella is any guide, her generation has picked up on the climate of ecological concern, but seems to have nebulous views about why they should be worried and what they should be doing in response. Ella is surely not alone in expressing particular concern for the fate of polar bears as the Arctic ice melts, a consequence of climate change that is much easier for a seven-year-old to grasp than are falling crop yields or desertification.

This combination of passion for the idea of environmentalism, and vagueness about details, is widespread among children. Publishers have spotted an opportunity and are rushing to publish environmental books for young readers. Some of these weave a subtle eco-message into a story, to instil a deeper understanding of natural processes and cycles; others are hectoring 'how-to' manuals that tell would-be planet-savers what to do.

Is That a Butterfly? is at the gentler end of the spectrum. A clueless bee and a well-informed snail watch and discuss the progress of tiny eggs as they develop into caterpillars and then butterflies, in a simple tale that explains the idea of life cycles. Readers can lift flaps to see what is going on underneath leaves, and the book even manages to explain caterpillars' use of prickly spines as a defence mechanism. *Why Should I Protect Nature?* goes a step further and explains why it is in humans' self-interest to respect and preserve the environment. Global warming is not mentioned. The much simpler message is far more likely to resonate with children. If you pick flowers and swat bees, for example, "we'd have no honey for breakfast"; and if you leave litter in the countryside it might harm the farm animals that produce milk and wool.

The green agenda is more prominent in *George Saves the World by Lunchtime* (4–7 years) (pictured overleaf), a jolly book that makes fun of the mundane nature of planet-saving while delivering some admirably pithy explanations: "When you throw things away, you are also throwing away the materials, the time and the energy it took to make them." The hapless George expects saving the world to be swashbuckling stuff. Instead he is taught to switch the lights off, do the recycling, take toys and clothes to a charity shop, buy locally produced strawberries and so on.

Children get their own version of the grown-up "my year as an eco-warrior" genre in *Spud Goes Green* (8–12 years) (pictured). Formatted as a diary, it starts with young Spud's New Year's resolution to "go green" with the help of his friend Adi, who provides the advice on such



GEORGE SAVES THE WORLD BY LUNCHTIME. EDEN PROJECT/RANDOM HOUSE

things as sealing draughty doors, recycling, living for a day without electricity, conserving water and recycling. The tone and layout are humorous, and pack in lots of silly factoids (some of them dubious) and craft projects, both of which appealed to Ella. "If you shout at a cup of water for eight years, it gets warm!" we both learnt. Ella particularly liked the idea of recycling her own paper.

101 Ways to Save the Earth (6–12 years) takes a more strait-laced approach in which a friendly whale lays out the basics of environmentalism, focusing on water, habitat, air, life and energy (which make the acronym WHALE). The book then suggests 101 ways to be green in everyday life, from putting a brick in the cistern to making your own compost. All these tips are labelled with whales sporting appropriate letters: W and E if a particular action saves water and energy, for example. Saving the Earth is defined quite broadly: there is a box on global warming, but readers might be forgiven for concluding that avoiding cosmetics tested on animals, and buying your cat a collar with a bell on it to give birds a sporting chance, are of equal importance. The author is naturalist and broadcaster David Bellamy, a climate-change sceptic who prefers readers to focus on conservation rather than carbon emissions; he provides a list of conservation organizations at the back of the book.

The tone of *Superkids* (8–12 years) is altogether more strident. As the jacket puts it: "Help! The Zombie Adults are messing up the world! They're killing our animals, wasting our water, destroying our trees, poisoning our food … Who can save the day? We need … the Super Kids." Saving the day does not simply mean protecting the environment. It means helping the homeless; cutting back on television, as "one in five children are affected by violence on TV" (adversely, one assumes); and not getting "caught up in the celebrity trap". Nuclear power is written off as a bad idea and genetically modified foods are said to contain "random bits of DNA, all stitched together", which will no doubt be news to their creators. These are things over which reasonable Zombie Adults can disagree.

Ella was most impressed by *An Inconvenient Truth* (7–18+ years), the book version of Al Gore's lecture and Oscar-winning film. With photographs, charts and not much text, it vividly explains the danger posed by climate change. "Part of some countries would go underwater … and there are lots more wildfires," she concluded. "It has lots of good pictures." As a keen student of hurricanes, tornadoes and other natural disasters, Ella was enthralled; there are even polar bears. Having made such a compelling case for action, the book's suggestions for what children can do seem rather feeble: switch to low-energy bulbs and try to use your car a bit less, essentially.

Perhaps that is inevitable. The underlying problem is that voluntary greenery by a few eco-conscious consumers in the wealthy West is not going to be enough. Voluntarism is a good way to practise for a carbon-restricted world, and can help to galvanize support for broader political action. Addressing the environmental challenges of the coming decades will require high-level political action by governments. A minority of consumers may choose to avoid incandescent lightbulbs and gas-guzzling cars, but governments can ban them outright. Ultimately, if Ella's generation is to save the world, it will be by voting for politicians who are prepared to impose tough restrictions on them. Perhaps they will be more inclined to vote for painful emissions cuts if they have grown up reading books like these and worrying about polar bears. ■

Tom Standage is business editor of *The Economist* in London, UK. His recent books include *A History of the World in Six Glasses*.

Is That a Butterfly
by Claire Llewellyn; Ant Parker (illus.)
Macmillan: £3.85

Why Should I Protect Nature?
by Jen Green; Mike Gordon (illus.)
Hodder Wayland: £5.99

George Saves the World by Lunchtime
by Jo Readman; Ley Honor Roberts (illus.)
Eden Project: £5.99

Spud Goes Green
by Giles Thaxton
Egmont: £4.99

101 Ways to Save the Earth
by David Bellamy; Penny Dann (illus.)
Frances Lincoln: £5.99

Superkids: 250 Incredible Ways for Kids to Save the Planet
by Sasha Norris; edited by Malcolm Tait; Rupert Davies (illus.)
Think: £5.99

An Inconvenient Truth: The Crisis of Global Warming
by Al Gore
Bloomsbury: £9.99

Hawking's fact and fiction

LUCY & STEPHEN HAWKING, GEORGE'S SECRET KEY TO THE UNIVERSE, RANDOM HOUSE

George's Secret Key to the Universe
By Lucy and Stephen Hawking
Doubleday: £12.99
Simon & Schuster: $17.99

George F. R. Ellis, with help from Ruby (10)

Stephen Hawking's book *A Brief History of Time* was a huge commercial success. Its achievements in bringing difficult scientific ideas to a wide audience are not so clear. Now the distinguished physicist has teamed up with his daughter Lucy to produce a children's book designed to communicate contemporary physics. Will it capture the attention of young minds and teach them some real science? Or will it be boring and over the heads of the prospective readers?

George's Secret Key to the Universe is an adventure story complete with villains and hero and is illustrated with enjoyable line-drawings. It involves a lost pig, a humorously portrayed intelligent computer, school bullies and a trip through the Solar System. Didactic discussions on aspects of modern physics, such as supernova explosions and black-hole physics are hung on this set-up. There are also fact boxes on physics and astronomy, and some photographs of astronomical phenomena: planets, comets, galaxies and so on. Overall, the book is a serious effort to convey facts and ideas in present day astronomy and astrophysics, within a science-fiction adventure story.

The mixture is great. Children love facts and adventure stories. The combination will catch their interest and keep them occupied for hours. After ten minutes of leafing through the book, my granddaughter Ruby was deeply absorbed and I had to promise to bring it back for her to read after I had completed my review. Like any educational tool, it will succeed for some and not for others. I reckon there should be more of the former.

I have two small quibbles. First, there is a bit of a disjunction between the science and the science-fiction parts that could confuse: is the kind of space travel envisaged in the narrative compatible with the hard science in the science sections? I suspect not. Second, I find the cover garish. But my youthful consultant loved it; so who am I to query the taste of youth? ■

George F. R. Ellis is professor of Mathematics at University of Cape Town, Rondebosch 7701, Cape Town, South Africa.

Stones, bones and stories

Henry Gee, with help from Phoebe (9) and Rachel (7)

The 'life-as-grand-narrative' school of children's books, such as Spinar and Burian's spectacular *Life Before Man*, first published in 1972, tell the history of life as 'Manifest Destiny', in which isolated fossil remains are seen as parts of a preordained jigsaw.

Books like that have an undeniable appeal — I loved them as a child. The determinedly old-fashioned *Life Story*, published in the late 1980s, is very much part of that tradition. "This book is captivating," says Phoebe, aged nine, "with its beautiful illustrations and words. I loved it!"

Phoebe is a chip off the old block, to whom the grand-narrative theme is an easy sell. The knack is to entice into the contemplation of life's splendid drama those children who might otherwise not have considered it — and, hopefully, to keep them there, enthralling them sufficiently that they spurn opportunities later on to become more interested in cell signalling or real estate.

How can this be done? Rachel is far less interested in the idea of dinosaurs as living animals than is her sister, but has a fine eye for a fossil in the field. For her, shapes and colours are as important as concept, and she enjoyed Neal Layton's calculatedly anarchic *The Story of Everything*, an elaborate pop-up book that looks like how *Life Story* would have turned out, had it been written by a mildly scatological graffiti artist.

In other words, what you need is a gimmick. *The Pebble In My Pocket* tries this through the valid, somewhat earnest mechanism of following the career of a pebble from its origins as volcanic lava until it ends up being picked up by a little girl 480 million years later: the history of life being told from the point of view of a pebble. The problem is that pebbles make unsympathetic narrators.

Ask Dr K Fisher About Dinosaurs is the other extreme — a scrapbook-like presentation of a collection of dinosaur problems solved by a prehistoric agony aunt. A juvenile *T. rex*, worried that his teeth are falling out, is consoled to learn that this is normal, and his teeth will be replaced. The children enjoyed the presentation, but it hardly lingered in the mind.

Fractionally more successful was *Prehistoric Actual Size*, in which parts of prehistoric animals — a tooth or a claw — are represented at their actual size. This is a clever idea, one of the thrills I get when seeing a museum specimen for the first time is realizing that the real thing is so much smaller (or larger) than I had imagined. But like many gimmicks, it's a one-trick pony.

The gimmick that works for everyone is to tell the history of life as a human story, in which real people are measured up against geological time, to give it scale. So the big hits *chez* Gee were *The Fossil Girl* and *Stone Girl Bone Girl*, two books about Mary Anning (1799–1847), the young woman who excavated many important specimens of early Jurassic marine life from the cliffs at Lyme Regis on the south coast of

Life Story
Eric Maddern & Leo Duff (illus.)
Frances Lincoln: £5.99
The Story of Everything
by Neal Layton
Hodder: £12.99
The Pebble In My Pocket
by Meredith Hooper & Chris Coady (illus.)
Frances Lincoln: £5.99
Ask Dr K Fisher about Dinosaurs
by Claire Llewellyn & Kate Sheppard (illus.)
Kingfisher: £7.99
Prehistoric Actual Size
by Steve Jenkins
Frances Lincoln: £11.99
The Fossil Girl
by Catherine Brighton
Frances Lincoln: £5.99/$7.95
Stone Girl Bone Girl
by Laurence Anholt & Sheila Moxley (illus.)
Frances Lincoln: £5.99/$7.95
The Human Story
by Charles Lockwood
Natural History Museum: £9.99

England. Both books are fictionalized and Mary Anning becomes every bit as mythical a character as Red Riding Hood. They found immediate acceptance with Rachel, who gravitates more to princesses than plesiosaurs. *Stone Girl Bone Girl* looks and feels indistinguishable from any bedtime story. This tried-and-tested formula still had enough facts about prehistoric life to satisfy Phoebe.

The Fossil Girl goes one better and does the childhood of Mary Anning as a really sharp graphic novel. This got full marks from Gees of all ages. Graphic novels could be the way to go: palaeoanthropology through the eyes of a six-packed Louis Leakey in the style of Judge Dredd? That has got to be better than *The Human Story*, the Natural History Museum's latest authoritative, uncool communiqué, brim-full of facts and threadbare on fun. Rachel says she might like such things, if they didn't have so many long words. ■

Henry Gee is a senior biological sciences editor at *Nature* and author of *Jacob's Ladder: The History of the Human Genome*.

Star tales

Mark Brake

"If you want your children to be intelligent," Albert Einstein said, "read them fairy tales. If you want them to be very intelligent, read them more fairy tales." Stories are how we make sense of the world, they help shape what we see, do and dream. Children who have difficulties focusing in class will sit spellbound by a narrative.

There are a lot of studies on the power of stories; psychologists refer to information presented in story form as 'psychologically privileged'. Our brains, it seems, are especially attentive and responsive to information conveyed in a narrative. Stories greatly aid recall. They provide a meaningful structure — hooks upon which to hang new knowledge.

All this is a gift, you would have hoped, to communicators of science. Yet too many authors of children's science books pick up on one of the key strands of storytelling without taking the creative step of bringing them all together. And what greater story than space and the origin of our planet?

Narrative structures make more sense when causal relationships are clear, and causality is key in Eric Maddern's *Earth Story* (8–13 years). This is a dramatic and nicely illustrated account of the origins of Earth. Maddern takes the reader on a journey from the enormous bang at the beginning of the universe through to the very first forms of life on our planet. But the book suffers from an otherwise limited and impersonal approach, with little suspense.

The use of complications and challenges in a narrative help to create a problem-solving scenario that involves the reader. *Stardust from Space* by Monica Grady (8–12 years) attempts to tell the story of stardust, and how it made our Solar System. But it is so replete and unrelenting in its presentation of data, one can almost hear Thomas Gradgrind, Charles Dickens's heartless utilitarian, screaming, "Fact, fact, fact!"

Strong, interesting characters are also essential to good stories. As F. Scott Fitzgerald went so far as to suggest, "action is character". Jacqueline Mitton understands this very well. Her books *Zoo in the Sky: A Book of Animal Constellations* and *Kingdom of the Sun: A Book of The Planets* (7–12 years) (pictured below), imbue the Sun, plants and constellations with character. Lavishly illustrated by Christina Balit, Mitton's stories clearly exhibit an understanding of the power of character and narrative for newly confident readers. Mitton does even better with *Galileo: Scientist and Star Gazer* (9–16 years). Here is the excitement of the discovery and exploration of science as unfolding narrative. A wonderful way of stimulating aspirant cosmologists.

A book that manages to bring together many of the key components of successful storytelling is David Donohue's *Moon Man* (8–14 years). This is a gripping and inventive account of young Walter Speazlebud's quest to find the truth about the 1969 Moon landings: fact or fiction.

Walter uses his power of Noitanigami — 'imagination' to those who sadly lack the ability to spell, talk and travel backwards. He travels back in time to 1969 to prove that the Moon landing happened. Granted, the Moon's surface does look a little like the Nevada desert. Granted too that Neil Armstrong botched his 'lines'. Walter's grandad said it happened, and even if he is getting confused in the head, he must be right. Walter proves this by mastering his gift of Noitanigami. His next task: to take his grandfather back in time, too, and rid him of his Alzheimer's disease. Marvellous. ■

Mark Brake holds a chair in science communication at the University of Glamorgan, UK and is co-author of *Different Engines: How Science Drives Fiction and Fiction Drives Science*.

Earth Story
by Eric Maddern & Leo Duff (illus.)
Frances Lincoln: £5.99

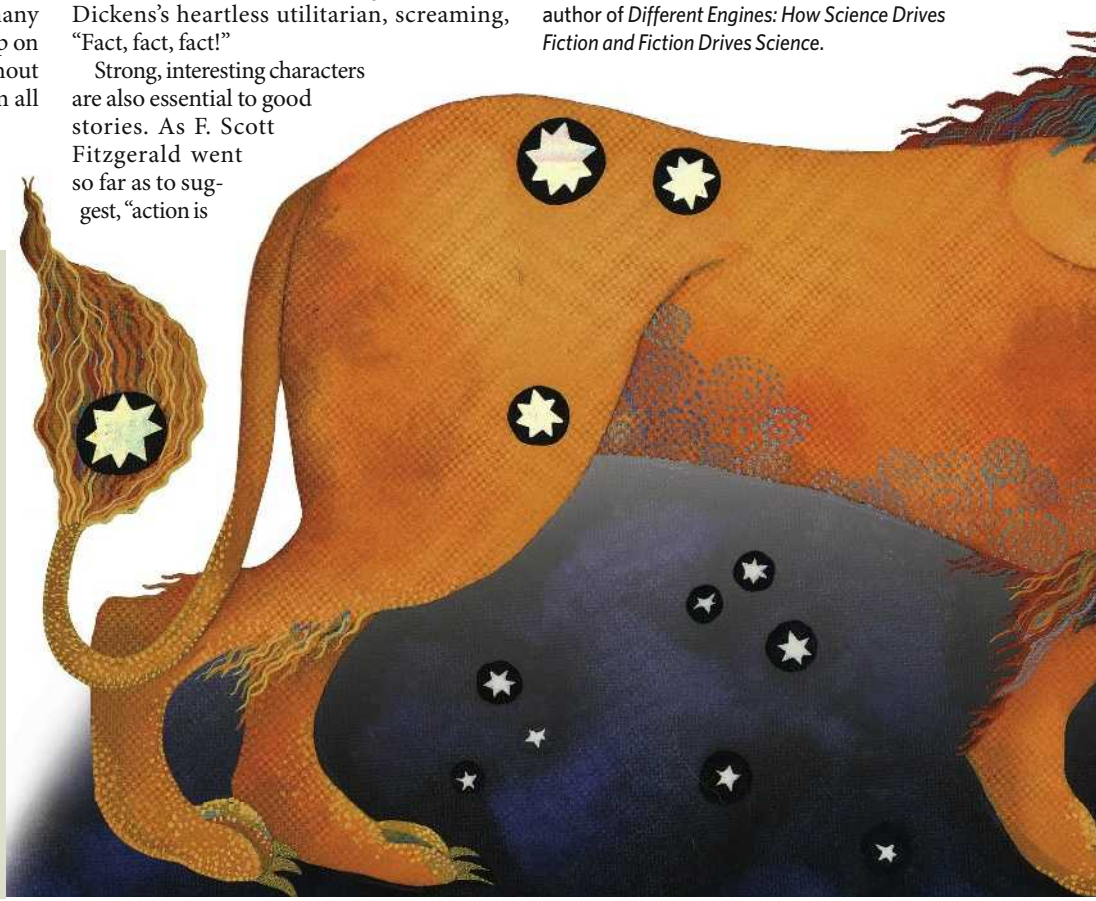
Stardust from Space
by Monica Grady & Lucia deLeiris (illus.)
Frances Lincoln: £11.99

Zoo in the Sky
by Jacqueline Mitton & Christina Balit (illus.)
National Geographic: $7.95

Kingdom of the Sun
by Jacqueline Mitton & Christina Balit (illus.)
Frances Lincoln: £6.99
National Geographic: $16.95

Galileo: Scientist and Star Gazer
by Jacqueline Mitton & Gerry Ball (illus.)
Oxford University Press: £4.99

Moon Man
by David Donohue
Egmont Books: £4.99



To bodily go ...

Ian Jones

In the film *Fantastic Voyage* (1966), a group of doctors are miniaturized and sent into the body of a defecting scientist to clear an untreatable blood clot. They witness close up processes at the heart of life — the oxygenation of a red blood cell, an electrical impulse in the brain. Journeying around the body is also the theme of *Inside You* (8–14 years), a book and accompanying CD-ROM. A futuristic 'nanocam' patrols the body providing "breaking news from the front lines of the body battlefield". It shows what happens when we get stung by a bee, vomit, squeeze a spot and urinate.

It is a laudable attempt to convey science to children aged 8–14 years. The implicit assumption is that exposure to the beauty of science will encourage the young to find out more and learn new things.

Yet the landscape of science education is more complex than is sometimes appreciated. It has been tempting to see the promulgation of science as a linear causal process in which specialist knowledge is transmitted to lay audiences who are blank slates that learn and become positively disposed to science.

This view is now seen as a caricature. There are a multitude of stakeholders with different, and not necessarily compatible, agendas: scientists keen to share their enthusiasm; advocates promoting positive attitudes to science; public-engagement professionals stimulating dialogue and debate; politicians concerned with national competitiveness; and audiences that are active interpreters of the information they receive.

Moreover, it is not just facts but the nature and process of science that need to be communicated. UK schools now include a course on 'How science works'. Arguably, it is better to know the principles of a randomized controlled trial than the names of all the vertebrae.

Mediators end up being pulled in different directions. Take science centres — their funders may have clear ideas about what they should promote, and they have to attract visitors to survive. Hence the proliferation of exhibitions on *Star Wars* and *The Hitchhikers Guide*, and the accompanying cries of dumbing down. Publishing is little different. It is a commercial activity and will survive only if it can make a profit.

Where does *Inside You* fit? The digital world provides great opportunities for communicating dynamic processes, and biology is nothing if not dynamic. The animations on the CD-ROM succeed admirably in bringing body processes to life, even if the abstraction at times borders on the psychedelic excesses of *Fantastic Voyage*.

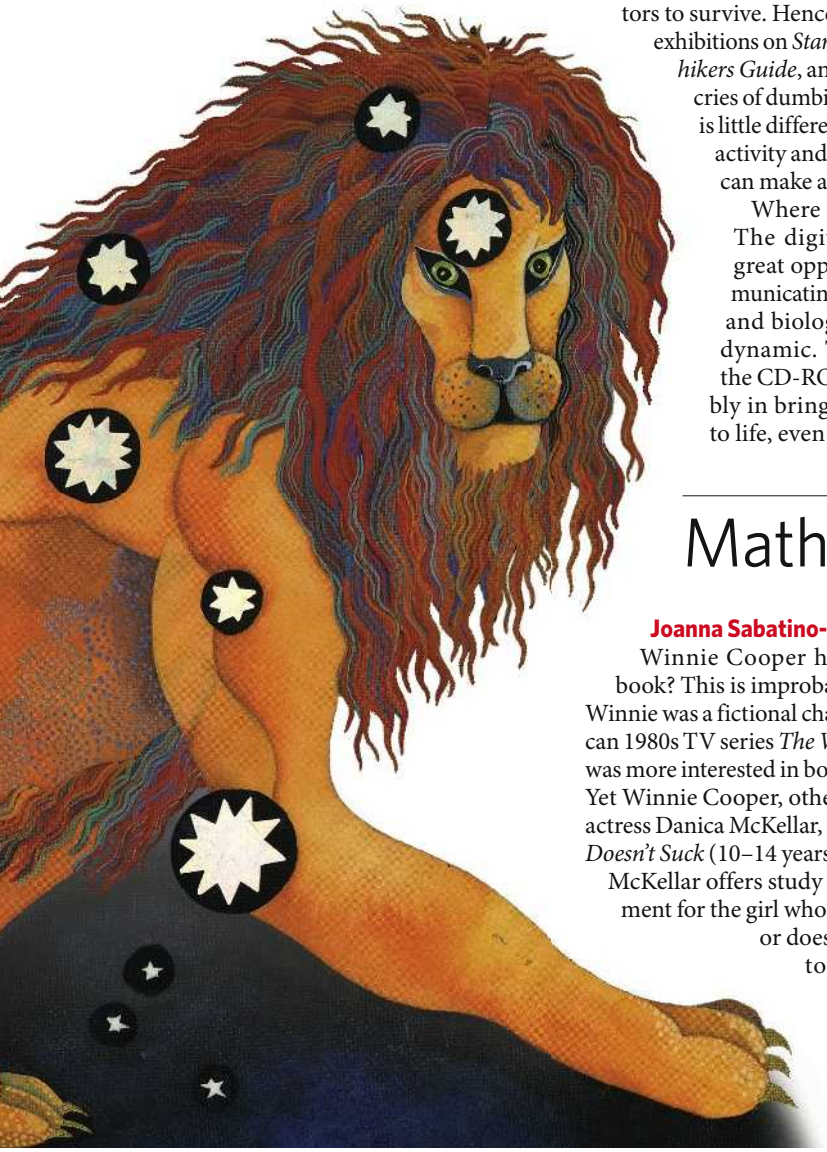
Translated to paper, though, some images take on a strange quality, caught between vague forms and elaborate, confusing textures. The text is mostly clear and simple, though sometimes quite technical. In a book that uses 'breathes in' instead of inhales, it is odd to see 'conjugation', 'scolex', 'proglottids' and 'protists'. If Stephen Hawking can avoid equations, can't biologists avoid unnecessary jargon?

Despite its high-tech premise, there is something old-fashioned about *Inside You* — the subject matter is practically *Fantastic Voyage* era. Hardly any recent developments in medical science are included. Readers get no sense of the innate and acquired immune responses, or the revolution in genomics. The notion of microbial communities is barely touched upon — every microbe is an enemy out to do us harm. If you are hoping to find out something about science or scientists, forget it. The book's fascination with the uglier side of the working body will no doubt appeal to its target audience. As a visually striking package it will engage children and they may well learn some facts.

The rumour mill suggests that *Fantastic Voyage* is being remade. Presumably, it will be another computer-generated spectacular. It may even be educational and tell us something about science — or is that asking too much? ■

Ian Jones is director of Isinglass Consultancy, Worldwide House, 22 Stephenson Way, London NW1 2HD, UK.

Inside You: How Your Body Makes it Through Every Day
by Mark Hamilton
Dorling Kindersley: £12.99



ZOO IN THE SKY, FRANCES LINCOLN PUBLISHERS

Mathematics not shopping

Joanna Sabatino-Hernandez

Winnie Cooper has written a maths book? This is improbable for two reasons. Winnie was a fictional character in the American 1980s TV series *The Wonder Years* and she was more interested in boys than in education. Yet Winnie Cooper, otherwise known as the actress Danica McKellar, is the author of *Math Doesn't Suck* (10–14 years).

McKellar offers study tips and encouragement for the girl who just isn't into maths, or doesn't see how it relates to her life. She also writes for that brainy girl who needs reassuring that being great with logic and numbers doesn't mean you are nerdy. Like a teen magazine for the mathematics classroom, *Math Doesn't Suck* empowers young girls with a funny, light and interesting tone.

As a middle-school maths teacher and mother, I see the daily struggle of girls who say, as the talking Barbie doll of the 1990s used to, "maths is hard, let's go shopping!". McKellar speaks to those kids with chapters such as, 'You Can Never Have Too Many Shoes' to teach multiples, and 'Is Your Sister Trying to Cheat You out of Your Fair Share?' to explain how to compare and convert fractional slices of pizza. Each chapter reinforces a single topic, from adding basic fractions to solving pre-algebra word problems.

She adds maths horoscopes, tips, quizzes ('Are you a Mathophobe?') and testimonials from girls who used to think that maths "sucked" and who now love the subject. There is even a

troubleshooting guide that includes web pages, help and other extras, and a linked website with more resources. McKellar also offers alternative strategies for multiple learning styles. Although this book can't replace a great teacher, it certainly helps support school instruction.

Math Doesn't Suck connects the ideas taught in the classroom to a young teen's daily life. My favourite example is how to find the 'Greatest Crush Factor', or GCF. She compares the factors that made her like her old "crush" with what makes her like her new crush. She lists the factors, circling what they have in common, identifying the greatest thing they have in common — *et voilà*, GCF.

Is it sexist to do whatever it takes to make sure that a girl likes maths early on in her education in order to pursue it as a career, or even to simply like it recreationally? I don't think so. Positive female role models such as McKellar — who embrace mathematics as a part of everyday life, and are still funny and cool — let a key audience know that it's okay to be great at mathematics.

Today's teenage girl may have no idea who Winnie Cooper is, but Winnie knows her! ■

Joanna S. Sabatino-Hernandez is an algebra and geometry teacher at Cabin John Middle School, 10701 Gainsborough Road, Potomac, Maryland 20854, USA.

Math Doesn't Suck
by Danica McKellar
Hudson Street Press: $23.95

The scene is set

Glenn Murphy

With the growth of the popular-science genre, science-related books for children have become more numerous and varied than ever before. Yet most kids would happily choose TV or MySpace over the average non-fiction text. So pitching the tone and content correctly is paramount. Authors have to do more than avoid the obvious pitfalls of technical language. They have to provide a solid context for every bit of content. If it isn't relevant to children or they can't recognize the familiar, why should they care?

Phillip Ardagh always succeeds at putting science and technology in context for children. His books are fun, engaging, larger than life and spilling over with interesting facts and anecdotes on every subject. *Wow! Inventions That Changed the World* (10–16 years) covers major breakthroughs in the history of science and technology.

Ardagh sticks to machines and vehicles rather than concepts or processes. Each chapter begins with a scene from an invention's history, such as the first phone call in Bell's laboratory: "Acid from the battery spills on to his trousers and Bell leaps to his feet. 'Mr Watson! Come here! I want you!' he cries … he has just made the world's very first telephone call." Ardagh then launches into the background and historical setting. From wheels to cars, mining carts to steam trains, he bobs lightly along, adding facts, figures and anecdotes along the way. The main themes — technology in our everyday lives and appreciating inventions in contrast to what went before — come through loud and clear. And like all the best educational texts, young readers remain unaware of how much they're really learning.

Books on animals have a head start with children. *Natural History Museum Animal Records* (6–14 years) (pictured) by Mark Carwardine is a compendium of animal facts (biggest, smallest, fastest, laziest, most dangerous, and so on).

NATURAL HISTORY MUSEUM ANIMAL RECORDS THE NATURAL HISTORY MUSEUM, LONDON/S. FRINK/DIGITAL VISION/GETTY

Most chapters are divided into linnaean orders. The bird chapter, in contrast, is sectioned into physiological and behavioural categories (such as talking birds, birds in flight). This neatly emphasizes the diversity of what to most children is a physically homogeneous group.

The book frames many 'record-breaking' animal characteristics as evolutionary adaptations, and features regular nods towards the relevance and importance of biodiversity and conservation. Like the arrangement, the tone is somewhat encyclopaedic, with a fair amount of jargon — making me wonder at times if this inspiring visual reference book is actually aimed at kids or at a more general family audience. Nonetheless, I would have loved a book like this as a child.

Conversely, *Actual Size* (4–8 years) by Steve Jenkins keeps it simple for younger readers, and features a collection of artistic collages of animals or parts of animals. Each is printed to scale across a double-page spread. These pictures are worth the proverbial 1,000 words of description. Accompanied by facts, statistics and dimensions — "A 60–centimetre-long tongue! This must be a giant anteater snacking on its favourite food, termites." — the illustrations are rounded off with a few pages of text about each animal. The idea is quite simple; the effect is wonderful.

Another way of contextualizing science is to relate it to the question 'what do you want to be when you grow up?'. *How To Be A Brain Surgeon* (8–12 years) by Amanda Li is part of a new series of books that does exactly this. You can just imagine a pushy parent buying it for their child in the hope of creating an aspiring 12-year-old neurosurgeon. Happily, the remit is wider than the title implies — exploring the history of medical science, technology and practice more generally before focusing on the brain and neurosurgery. The tone drifts into the formal, passive voice of the academic curator at times, but generally clips along well. The book offers some genuine insights and background for kids where otherwise there is none, and it avoids being too preachy or sugar-coating the themes.

These books vary in approach, tone and target age range. In providing children with a context for science and technology in their lives, all of them have something to offer younger readers. The central importance of context is easily forgotten in the race to produce 'educational' books aimed at improving the 'scientific literacy' of children and young adults. Education begins with engaging interest. Without that, most attempts at imparting information — whether formal or informal — will either fail or be actively counterproductive. So if you're buying science-related books to encourage your offspring, choose carefully. The wrong books will put them off science completely. The right ones can inspire them for life. ■

Glenn Murphy is a US science writer and is the author of *Why Is Snot Green?*

Wow! Inventions That Changed the World
by Phillip Ardagh
Macmillan: £3.99
Natural History Museum Animal Records
by Mark Carwardine
Natural History Museum: £20
Actual Size
by Steve Jenkins
Francis Lincoln: £9.99
How To Be A Brain Surgeon
by Amanda Li
Macmillan: £3.99

NEWS & VIEWS

PHOTONICS

Rogue waves surface in light

Dong-Il Yeom and Benjamin J. Eggleton

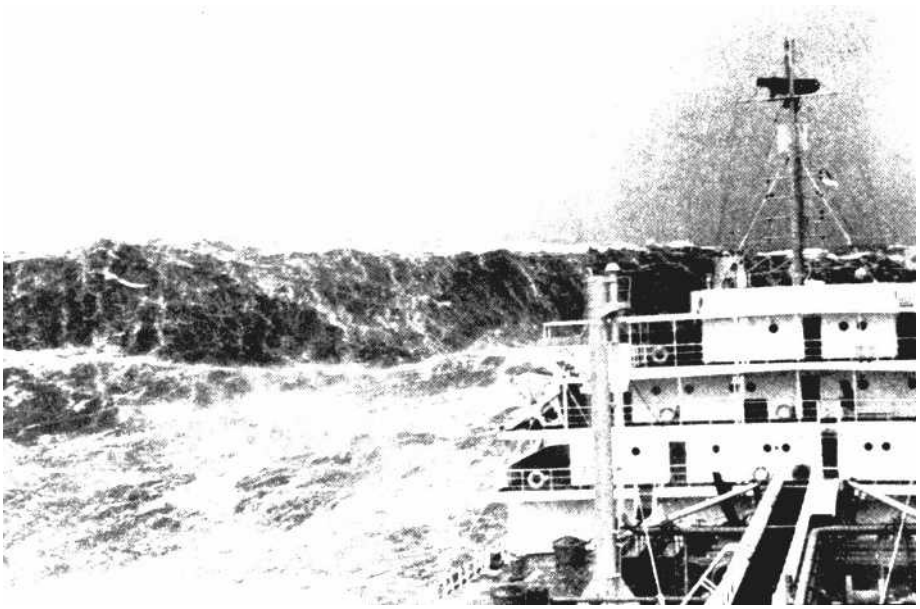
How do the freak waves that haunt seafarers' nightmares arise? We don't know, is the short answer — but the discovery of a similar phenomenon in optical waves might assist in getting to the bottom of the mystery.

Oceanic rogue waves — monstrous sea waves that form spontaneously and can reach up to 30 metres in height[1,2] — have been held responsible for marine misfortunes ranging from the sudden sinking of seagoing ships to damage to oil platforms. They are not just the stuff of maritime folklore (Fig. 1): recent satellite data have revealed that extraordinarily large waves are in fact far more frequent than statistics would predict. The phenomenon remains poorly understood, and the difficulty of studying rogue waves under controlled conditions makes investigation highly problematic.

Until now, perhaps. On page 1054 of this issue[3], Solli *et al.* detail a series of experiments in a 'nonlinear' optical medium — a photonic crystal fibre — that could enhance our understanding of the physics underlying rogue-wave formation and propagation. The authors claim to have found an optical counterpart of the hydrodynamic rogue wave. It occurs in a system that can be studied in a controlled way using off-the-shelf components, and that delivers results that are readily comparable with well-established theoretical and numerical models.

Solli and colleagues' essential argument is that rogue waves, whether oceanic or optical, are associated with solitons. Generally speaking, solitons are particularly robust solitary wave packets that propagate in a dispersive medium (one in which a wave's speed depends on its wavelength) without becoming distorted. But in the 'noisy' environment of the ocean — where wave noise can result from any number of factors, such as a change in wind direction — or in a specially designed noisy optical system, solitons can also be generated through a process known as modulation instability. In particular, the authors show how this noise seeding, combined with highly nonlinear propagation through the optical fibre, causes the generation of occasional solitons that are very different from normal. These solitons appear as a dramatic intensity spike on top of a low-amplitude background — in other words, as rogue waves.

The authors[3] exploit a nonlinear phenomenon in an optical fibre, known as supercontinuum generation[4], to excite and then study optical rogue waves. In supercontinuum generation, which is a well-understood phenomenon[5], light initially in a narrow band of frequencies is dramatically converted into ultra-broadband light. The effect has been used for applications in many diverse fields — notably in laser-frequency metrology[6], for which half of the 2005 Nobel physics prize was awarded. The noise properties of the broadband spectra have also been investigated in detail, as they are pivotal in assessing potential applications. Research was initially concentrated on the mechanisms by which noise from very short input pulses is transferred to the output spectrum, but it has since spread to cover the full range of possible inputs, from femtosecond pulses right up to continuous light waves[7].



NOAA/NWS

Figure 1 | Science fact. Oceanic rogue waves are understandably rarely caught on camera. This instance, published in the Fall 1993 issue of *Mariners Weather Log*, was captured around the '100-fathom curve', where the continental shelf drops into the Atlantic abyssal plain in the Bay of Biscay — a known hotspot for extreme waves.

Until now, however, optical studies have measured the stability of a supercontinuum only indirectly, by measuring noise at radio frequencies or through averaged optical spectra. Solli and colleagues' method allows the statistics of the supercontinuum output spectra to be measured directly; the statistics measured are the distribution of single-pulse properties from a number of pulse trains. The authors use a novel wavelength-to-time transformation technique that temporally stretches a large number of ultrashort pulses by passing them through the dispersive medium. In this way, many random events generated by trains of regular input waves to which noise is added can be measured as they happen (Fig. 2a, overleaf).

The authors could thus show directly that supercontinuum generation intrinsically creates a small number of 'rogue' waves of far greater amplitude than the norm. These events are associated with certain solitons shifting towards longer wavelengths, and the effect appears distinctly in the spectrum as the pulse propagates along the fibre (Fig. 2b). The statistics of these optical rogue solitons are similar to those of oceanic rogue waves: plotting a histogram of the probability of encountering waves of different amplitudes results in a characteristic L-shape, with more high-amplitude events than would be expected

NATURE|Vol 450|13 December 2007

50 & 100 YEARS AGO

50 YEARS AGO

"Scientists in Society To-day", proposal of a toast of the Royal Society by the Right Hon. The Viscount Hailsham Q.C.

To-night I can be as bold as brass. Although not a scientist, I am at least an 'egghead' by conviction and, I hope, by practice, and I am addressing a society of scientists who are also, always by achievement and almost by definition, 'eggheads'. It is time we got together. 'Eggheads' of the world, unite! We have nothing to lose but our brains. A country neglects its 'eggheads' at its peril. For it is the 'egghead' who is the greatest realist. It is the 'egghead' who invents the *Sputnik*, not the captain of football, nor the winner of the sword of honour, nor the president of the Junior Common Room ... It is a formidable indictment of Western civilization and democracy that 'eggheadedness' is not valued at its proper worth.

From *Nature* 14 December 1957.

100 YEARS AGO

A telegram from Largs states that Lord Kelvin has not been well for more than a fortnight, and has been confined to his bed. His condition on Tuesday night had improved. [But worse news was to follow in the *Nature* issue of 19 December 1907, as will be reported in 100 Years Ago next week.]

ALSO:

A proposal made to the Public Control Committee of the London County Council by Signor D. Maggiora to apply the process of discharging cannon of special construction, known in Austria as weather shooting, "to prevent the formation of fog or to disperse it in the case it is already formed, and also to disperse and destroy all clouds, and to prevent rain, hailstorms, lightning, and thunder," has been under the consideration of the Council. It was referred to the director of the Meteorological Office for report ... As might be expected, Dr. Shaw's report ... is entirely unfavourable.

From *Nature* 12 December 1907.

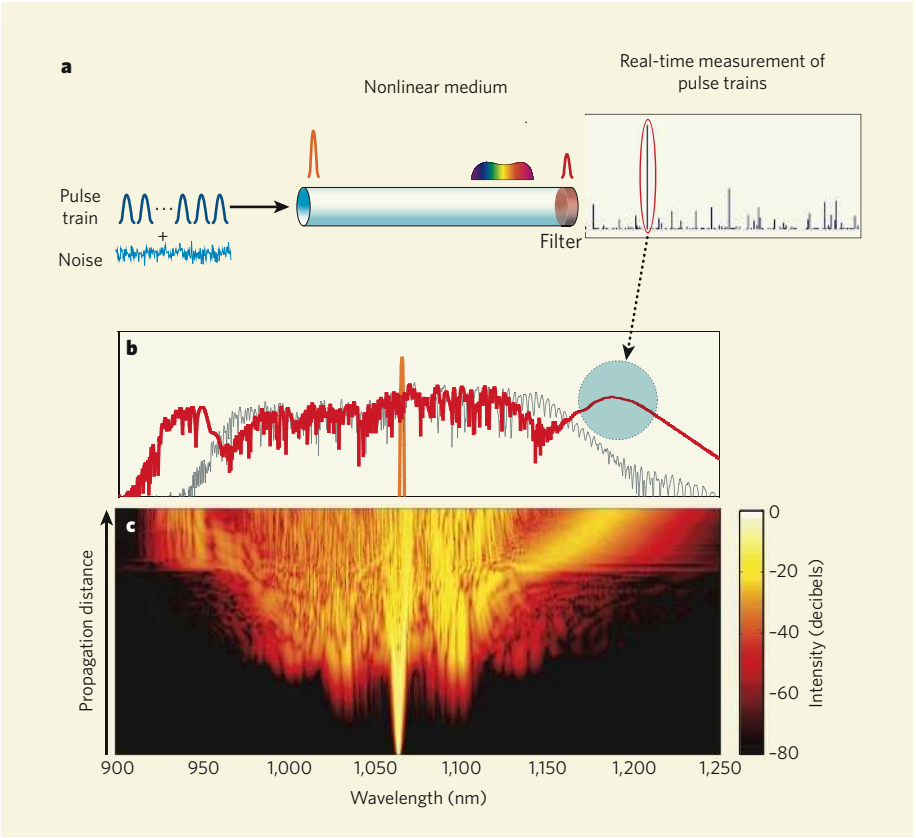


Figure 2 | Rogue generation. a, In Solli and colleagues' optical rogue-wave production[3], noise is added to a smooth wave pulse that is sent through a nonlinear medium (a photonic crystal fibre). In a process known as supercontinuum generation, this narrowband pulse is increased hugely in bandwidth. After spectral filtering, a small number of 'rogue wave' events of statistically abnormal amplitude are found in the real-time measurement of pulse trains. **b**, By following the evolution of supercontinuum generation along the propagation distance in the optical fibre, the authors show that the rogue events correspond to solitary wave packets (solitons) that are shifted to long wavelengths. The orange line represents the spectrum of the (narrowband) input pulse, the grey line that of a normal broadband output pulse, and the red line that of a rogue output pulse. **c**, From bottom to top, the evolution of the spectrum of rogue-wave intensity as the light propagates along the nonlinear medium.

from a conventional gaussian distribution. In addition, optical and oceanic waves both undergo dramatic fluctuations in intensity during their evolution.

Solli *et al.*[3] find supporting evidence for the soliton interpretation from numerical simulations, using the nonlinear Schrödinger equation — a way of representing wave propagation in a medium such as an optical fibre — to model supercontinuum generation. On the strength of these simulations, the authors propose an intimate connection between the initial amplification of input noise and the generation of rogue solitons.

So how do these optical findings help us to understand the generation of rogue ocean waves? Clearly, the noise and propagation environments are hardly identical. On the other hand, nonlinear propagation in optical systems is increasingly giving wider insight into areas as disparate as superfluidity and the science of self-similarity[8,9]. The direct experimental access to the statistics of optical rogue solitons, as well as the ease of 'managing' supercontinuum generation by modifying the optical-fibre geometry[10,11], means that rogue-wave physics is likely to join this list. The next intriguing stage will be to determine the precise degree to which the ideas elucidated by Solli *et al.*[3] transfer to the oceanic context. ■

Dong-Il Yeom and Benjamin J. Eggleton are in the Centre for Ultrahigh-Bandwidth Devices for Optical Systems (CUDOS), School of Physics, University of Sydney, New South Wales 2006, Australia.
e-mails: yeom@physics.usyd.edu.au; egg@physics.usyd.edu.au

1. Hopkin, M. *Nature* **430,** 492 (2004).
2. Perkins, S. *Science News Online* **170,** 328–329 (2006).
3. Solli, D. R., Ropers, C., Koonath, P. & Jalali, B. *Nature* **450,** 1054–1057 (2007).
4. Ranka, J. K., Windeler, R. S. & Stentz, A. J. *Opt. Lett.* **25,** 25–27 (2000).
5. Alfano, R. R. *Sci. Am.* **295** (6), 64–71 (2006).
6. Udem, Th., Holzwarth, R. & Hänsch, T. W. *Nature* **416,** 233–237 (2002).
7. Dudley, J. M., Genty, G. & Coen, S. *Rev. Mod. Phys.* **78,** 1135–1184 (2006).
8. Wan, W., Jia, S. & Fleischer, J. W. *Nature Phys.* **3,** 46–51 (2007).
9. Dudley, J. M., Finot, C., Richardson, D. J. & Millot, G. *Nature Phys.* **3,** 597–603 (2007).
10. Birks, T. A., Wadsworth, W. J. & Russell, P. St. J. *Opt. Lett.* **25,** 1415–1417 (2000).
11. Kutz, J. N., Lyngå, C. & Eggleton, B. J. *Opt. Express* **13,** 3989–3998 (2005).

MALARIA

Differential parasite drive

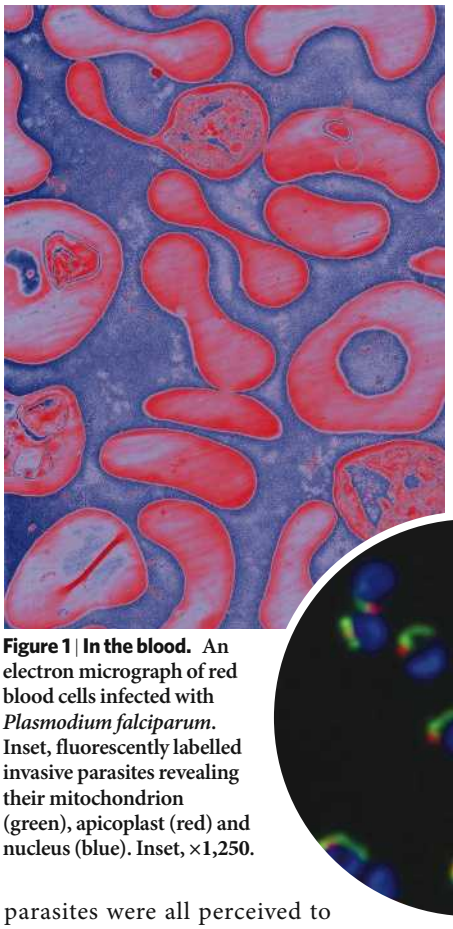
Giel G. van Dooren and Geoffrey I. McFadden

Our knowledge of the inner workings of malaria parasites comes largely from lab-based studies. But parasites growing in humans may have greater metabolic flexibility than those growing in Petri dishes.

Malaria parasites kill more than a million people every year. These minuscule organisms, belonging to the genus *Plasmodium*, ensconce themselves inside our red blood cells. They eat our oxygen-carrying haemoglobin protein, and sup on the rich supply of glucose in our blood plasma. Hidden from our immune system within our own cells, they multiply exponentially, inducing anaemia, acidity of the blood, low blood sugar, fluid build-up in the lungs, seizures and blockage of brain capillaries — complications that can kill a person within ten days of being infected by a malaria-carrying mosquito. Until now, we believed that malaria parasites burned the glucose they stole from our plasma using a simple and relatively inefficient process known as glycolysis. After all, why would a parasite bother to extract maximum energy from glucose when abundant free glucose is at hand?

On page 1091 of this issue, however, Daily and colleagues[1] show that, in some infections, the parasites behave as if they are starving, cranking up genes involved in energy-harvesting pathways to wring out the maximum burn from the proceeds of their parasitism. Switching on these genes could enable parasites to engage the tricarboxylic acid (TCA) cycle, the cellular motor that burns the leftover fuel from glycolysis to allow energy production to shift into top gear (see Fig. 4 of the paper[1] on page 1093). Intriguingly, the different parasite behaviours might correlate with different disease profiles, potentially explaining why different patients experience radically different symptoms during severe malaria infections.

Without a straightforward, easily accessible animal model for the deadliest malaria species, *Plasmodium falciparum*, lab work on the disease has been a difficult proposition. In 1976, a seminal paper[2] described a method of growing *P. falciparum* in Petri dishes of glucose-rich human blood with reduced oxygen levels. This method, essentially unchanged, is used in all modern malaria labs; it underpins all quests for a cure, whether they involve drug screens, genetic studies, genome sequencing, immunology, biochemistry or cell biology. Post-genomic studies have painstakingly mapped the expression levels of every gene and the quantities of each encoded protein across the orderly 48-hour part of the life cycle that the parasite executes in red blood cells[3–5]. These studies reinforced the view that genes encoding the proteins required for glycolysis are abundantly expressed during this part of the life cycle. The parasites were all perceived to be marching to a rigid, yet energy-profligate rhythm, apparently oblivious to the world around them — us.



PHOTOTAKE INC./ALAMY

G. G. VAN DOOREN/G. I. MCFADDEN

Figure 1 | In the blood. An electron micrograph of red blood cells infected with *Plasmodium falciparum*. Inset, fluorescently labelled invasive parasites revealing their mitochondrion (green), apicoplast (red) and nucleus (blue). Inset, ×1,250.

Studying pandas in a zoo is not the same as studying them in the wild. Likewise, these lab-based *in vitro* culture studies might not accurately reflect how the parasites behave in their natural environment. Hence the approach adopted by Daily *et al.*[1], who took blood from patients infected with *P. falciparum* in Senegal, and used microarray DNA chip technology to generate a parasite gene-expression profile for each patient. In many patients, parasite profiles were the same as those observed in lab-cultured parasites — the parasites seemed to be running on energy derived from glycolysis.

But parasites from two other groups of patients exhibited very different, and hitherto unseen, gene-expression profiles. In one group, the parasites had upregulated stress-response genes, probably to cope with host immune pressure. In the other group, they had downregulated the glycolysis genes normally switched on in lab-cultured parasites, but upregulated genes involved in alternative means of energy generation, a trait seen, for instance, in yeast cells starved of glucose. These results imply that there are physiological differences in the growth of parasite populations in different individuals. In other words, malaria parasites do not always grow in humans as they do in Petri dishes.

The changes in gene expression seen in 'starved' parasites are particularly curious. Two subcellular compartments (organelles) in the malaria parasite, the apicoplast and the mitochondrion, may hold the key to the apparent metabolic switching seen here. The apicoplast is a chloroplast-like organelle thought to have been originally photosynthetic but now retained for the biosynthesis of lipid building-blocks known as fatty acids[6]. Mitochondria harbour the enzymes of the TCA cycle as well as an energy-generating electron-transport chain that together finish the incineration of glucose and other substrates to increase energy yields. Both apicoplast fatty-acid synthesis and mitochondrial energy-production genes are dramatically upregulated in the 'starved' parasites[1].

What drives these differences in gene expression? Although apicoplast fatty-acid biosynthesis is essential for successful infection[7], its exact role is unclear because malaria parasites can scavenge fatty acids from their host. Upregulation of the apicoplast pathway for fatty-acid synthesis may suggest an increased need for fatty acids or a short supply of them from hosts. Similarly, upregulation of pathways involved in efficient mitochondrial energy generation implies either an increased need for energy or a reduced supply of glucose for glycolysis. A recent study[8] on laboratory-grown parasites concluded that mitochondria are not required for energy generation. But Daily and colleagues' discovery of the upregulation of genes involved in these mitochondrial energy-synthesizing pathways suggests that this may not always be the case.

The findings presented here raise various questions. What causes the different parasite gene-expression profiles? Do these profiles reflect distinct temporal stages of *in vivo* parasite development, or are they discrete snapshots of an intense battle between parasite and host? Are parasites initiating these differences, or are they merely reacting to cues from the host? Patient factors such as blood glucose levels, the amount of haemoglobin and the number of parasites in the blood do not seem to be linked to the starvation response. If there is an environmental cue, it is a subtle one.

One caveat in interpreting Daily and colleagues' results[1] is that gene upregulation doesn't always translate to metabolic upregulation; biochemical validation of actual metabolic switching is needed, and this will probably require elicitation of the starvation response in laboratory-grown parasites. At a more practical level, it will be important to understand whether different parasite gene-expression profiles are linked to the spectrum of disease experienced by patients with malaria, which, in turn, may point to more effective treatments. For example, drugs targeting fatty-acid biosynthesis and the mitochondrial electron-transport chain[8,9] should be especially effective against parasites in starvation mode.

We have come a long way in understanding malaria and its causes. But the findings presented by Daily *et al.*[1] show that we are just beginning to comprehend the complexity of the metabolic engine that drives these parasites. ■

Giel G. van Dooren is at the Center for Tropical and Emerging Global Diseases, University of Georgia, Athens, Georgia 30602, USA. Geoffrey I. McFadden is at the Plant Cell Biology Research Centre, School of Botany, University of Melbourne, Parkville, Victoria 3010, Australia.
e-mails: giel@uga.edu; gim@unimelb.edu.au

1. Daily, J. P. *et al. Nature* **450,** 1091–1095 (2007).
2. Trager, W. & Jensen, J. B. *Science* **193,** 673–675 (1976).
3. Bozdech, Z. *et al. PLoS Biol.* **1,** e5 (2003).
4. Le Roch, K. G. *et al. Science* **301,** 1503–1508 (2003).
5. Florens, L. *et al. Nature* **419,** 520–526 (2002).
6. Ralph, S. A. *et al. Nature Rev. Microbiol.* **2,** 203–216 (2004).
7. Waller, R. F. *et al. Proc. Natl Acad. Sci. USA* **95,** 12352–12357 (1998).
8. Painter, H. J., Morrisey, J. M., Mather, M. W. & Vaidya, A. B. *Nature* **446,** 88–91 (2007).
9. Goodman, C. D. & McFadden, G. I. *Curr. Drug Targets* **8,** 15–30 (2007).

ASTRONOMY

Dim view of past clashes

Curtis Struck

Simulations indicate that faint galaxies of a seemingly tranquil class were born in violent cosmic encounters. This would be good news for the prevailing model of how the Universe is constructed.

Giant low-surface-brightness galaxies, or GLSBs, are gentle giants of the cosmos. Their gaseous disks are up to 100 kiloparsecs across, several times the size of our Milky Way, yet rates of star formation within them are very low[1]. This makes them difficult to spot: the prototypical GLSB, Malin 1, was found only in 1986, despite being, at that time, the largest spiral galaxy ever seen[2,3]. But GLSBs are not just troublesome to observe: their existence is also a challenge for the prevailing 'cold dark matter' (CDM) model of the Universe's composition. Writing in *Monthly Notices of the Royal Astronomical Society*, Mapelli *et al.*[4] now propose an origin for these perplexing galaxies that might circumvent that difficulty.

The CDM model predicts that structure in the Universe is built up hierarchically from smaller units, powered by the gravitational force of 'dark matter' that cannot be seen at any wavelength. With the addition of a cosmological constant or other driver of accelerating cosmic expansion, the CDM model can account for a wide range of cosmological observations[5]. The impressive resolution of recent CDM-based simulations of the formation of structure in the Universe, such as the Millennium Simulation[6], is testament to the model's success.

But some problems remain. One is that CDM simulations tend to produce galaxy disks that are more compact and have less net angular momentum than those of real galaxies. With sufficient numerical resolution, and adequate treatment of feedback effects, the models do produce normal disks in some environments[7], and future models will possibly do better. But the existence of the hugely extended GLSB disks is an extreme challenge to this optimism. Mapelli *et al.*[4] first review several suggestions for the formation mechanism of GLSBs beyond standard CDM processes, and find them wanting. But they provide a new and rather surprising possible solution to the dilemma: that GLSBs are in fact the descendants of another cosmic curiosity, ring galaxies.

In the classic theory, the characteristic bright circle of a ring galaxy is produced when a secondary companion galaxy passes through the centre of a primary disk of stars[8] (Fig. 1). The stars and gas clouds in the primary disk are drawn inwards by the passing galaxy's gravity, but rebound when the companion moves away and the gravitational pull is reduced. The timescale for these effects increases with radius, so when particles farther in are on the rebound, those farther out are still falling in. The result is a circular wave of compression that propagates outwards, triggering the formation of spectacular clusters of bright stars as it passes. As these clusters flare up and slowly fade away again, they form a bejewelled ring that moves slowly outwards over hundreds of millions of years.

Many aspects of this standard theory of ring galaxies have been confirmed by

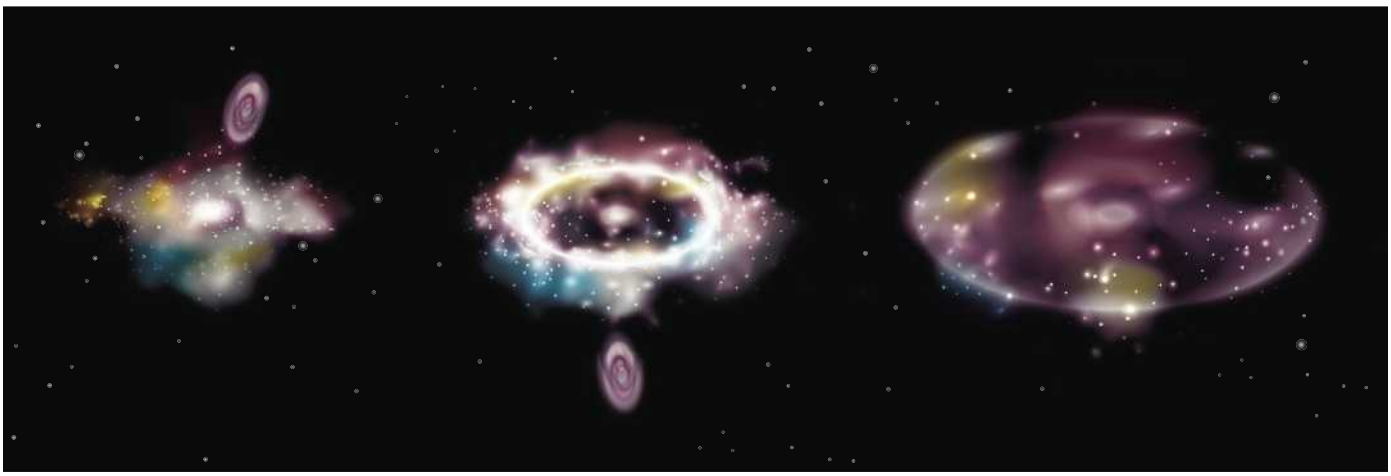
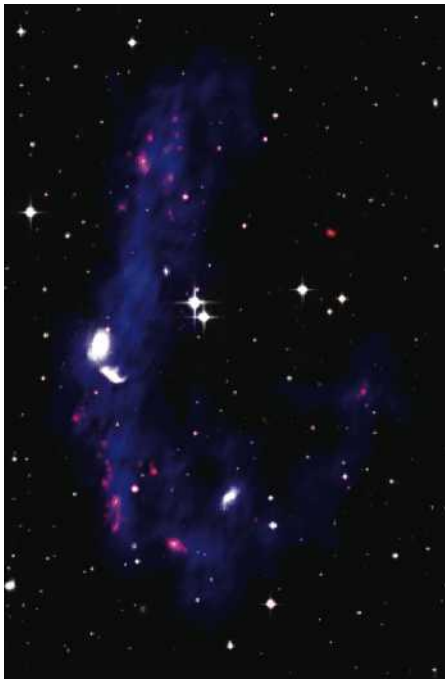


Figure 1 | The twilight of the ring. According to Mapelli and colleagues' simulations[4], the anomalous giant low-surface-brightness galaxies (GLSBs), whose formation is such a problem for the cold-dark-matter model of the Universe's structure, are the embers of cataclysmic galactic collisions. A smaller companion galaxy passes through the centre of a larger primary galaxy, generating a shock wave that ripples outwards, igniting gas to form stars — a phenomenon that we see as a circular ring galaxy. If the collision is particularly violent, the ring becomes hyper-extended and its intensity diminished. What we observe (right) is the distended, dim disk of stars left behind — the GLSB.



P.-A. DUC/CEA-CNRS/NRAO/AUI/NSF/NASA.

Figure 2 | Collision remnant. A multi-wavelength image of the hyper-extended ring galaxy NGC 5291.

observation[9], especially of systems where the companion is somewhat less massive than the primary disk, and is in or near a gravitationally bound orbit. Recent extensions have taken into account observations of several extreme processes, among them high-velocity collisions between galaxies that fall together from distant points, but that are both bound to a greater structure such as a galaxy group or cluster. The interacting system NGC 5291 (ref. 10) is a prominent example of the hyper-extended rings that can result (Fig. 2).

In large rings, the gas density is likely to be low, so the star formation rate will also be low. Mapelli and colleagues' central idea is that such a weak ring will generally be so dim as to be barely visible, and that the huge disk that it leaves behind it as it propagates outwards will look remarkably like a GLSB. Moreover, in the time (a billion years or so) that it takes the material to travel out to such large distances, the high-speed companion will in many cases have moved off the immediate scene. It is therefore hardly surprising if we see no obvious evidence of the past fracas.

The authors[4] use extensive data on four GLSBs to select, from a grid of computer simulations with a range of initial collision parameters, those models, and the times during their evolution, that are most like the observed systems. In all four cases, a model is found in which the surface-brightness profiles, optical colours, and the gas distributions and kinematics of the simulated galaxy all agree well with observations. In three out of four cases, possible companions are also in view.

The obvious consequence of the hypothesis is that GLSBs are disturbed bodies, and not the stable, quiescent galaxy disks that they had been assumed to be. This means that hierarchical galaxy-formation theory within a CDM model is not required to yield GLSBs as final products: the challenge they represent to the theory vanishes. In fact, the best-fit models of Mapelli *et al.* all include 'haloes' of CDM surrounding the colliding galaxies.

Sound as it might like a fractured fairy tale or a Wagner opera gone wrong, the evidence seems to suggest that cosmic gentle giants spring from bejewelled rings. But mysteries remain, such as how smaller, gas-rich low-surface-brightness galaxies form, and how they persist so quietly until they are lit up by a collision. With rapid advances in our understanding of galaxy formation, the prospects are bright for a speedy resolution. ■

Curtis Struck is in the Department of Physics and Astronomy, Iowa State University, Ames, Iowa 50011, USA.
e-mail: curt@iastate.edu

1. Pickering, T. E., Impey, C. D., van Gorkom, J. H. & Bothun, G. D. *Astron. J.* **114,** 1858–1882 (1997).
2. Impey, C. & Bothun, G. *Astrophys. J.* **341,** 89–104 (1989).
3. Edmunds, M. G. *Nature* **341,** 105–106 (1989).
4. Mapelli, M. *et al. Mon. Not. R. Astron. Soc.* (in the press); preprint at www.arxiv.org/abs/0710.5354 (2007).
5. Springel, V., Frenk, C. S. & White, S. D. M. *Nature* **440,** 1137–1144 (2006).
6. Springel, V. *et al. Nature* **435,** 629–636 (2005).
7. Governato, F. *et al. Mon. Not. R. Astron. Soc.* **374,** 1479–1494 (2007).
8. Lynds, R. & Toomre, A. *Astrophys. J.* **209,** 382–388 (1976).
9. Appleton, P. N. & Struck-Marcell, C. *Fund. Cosmic Phys.* **16,** 111–220 (1996).
10. Bournaud, F. *et al. Science* **316,** 1166–1169 (2007).

STRUCTURAL BIOLOGY

Ion pumps made crystal clear

David C. Gadsby

The function of every cell in our bodies depends on the work of proteins known as ion pumps. Several new crystal structures cast fresh light on how three different pumps deal with their distinct cargoes of ions.

Ion pumps toil tirelessly in cells throughout all kingdoms of life, transporting ions across membranes. To investigate the workings of these microscopic machines, X-ray crystal structures of a calcium ion pump known as SERCA have been determined[1–5]. But although those structures depict SERCA in several conformations, none of them caught the pump in the act of releasing its cargo of ions. Moreover, nagging questions remained about how much SERCA might differ from other, genetically related ion pumps — such as those that transport ions of different sizes and charges from calcium, or that require additional protein subunits. In this issue, three papers[6–8] from the same group go a long way towards addressing those concerns by describing the first atomic structures of a SERCA pump with its ion pathway open[6] and of two related proteins — a sodium–potassium pump[7] and a proton pump[8].

The three pumps described in these papers belong to a family known as phosphorylated-type (P-type) pumps, named after the phosphate group whose addition and removal controls their activity. P-type pumps inhabit all our cells and are essential for life. Without sodium–potassium pumps, many vital functions would fail. For example, there would be no electrical signals in our brains or hearts, or in any nerves or muscles; and without SERCA pumps, there would be no muscle contraction.

Not surprisingly, P-type pumps are hot targets for therapeutics — for example, digoxin (a treatment for heart problems) targets sodium–potassium pumps, and the latest antacids act on the proton–potassium pumps in our stomachs. So the stakes are high — determining the structure and mechanism of each P-type pump is crucial for further drug discovery.

P-type pumps reside either in the surface membranes of cells or in the membranes of intracellular organelles such as the endoplasmic or sarcoplasmic reticulum. In all cases, one end of the pump opens to the cytoplasm and the other end opens either to the outside of the cell or to the interior (lumen) of the organelle. The pumps adopt two main conformations[9], known as E1 and E2 (Fig. 1, overleaf). The ion-binding sites are found deep inside the region of the pump that crosses the membrane; in E1, these sites are accessible to ions in the cytoplasm. Ion binding promotes the phosphorylation of the pump, in which a phosphate group is added to a single amino-acid residue. The source of the phosphate is an ATP molecule; a side product (ADP) is formed that briefly remains associated with the pump. In the resulting E1P state, the bound ions are occluded — they are inaccessible from either side of the membrane. The pump then releases the ADP and relaxes to the E2P conformation, whereupon a pathway opens to the extra-cytoplasmic side, allowing the ions to escape.

Transport in the reverse direction begins when ions from the cell exterior or the sarcoplasmic reticulum bind to the exposed binding sites in the E2P state, triggering dephosphorylation of the pump. This yields another state with occluded ions, E2. The pump then relaxes back to the E1 state, reopening the ion pathway

to the cell interior and releasing the counter-transported ions to the cytoplasm.

The cycle is strictly controlled so that access to the ion-binding sites alternates between the two sides of the membrane[10], with the ions becoming temporarily occluded after each ion-binding event. Evolution has tailored this mechanism to transport ions against the prevailing ion gradient while avoiding ion leaks in the opposite direction. Because the occluded states are relatively stable, it is these conformations that predominate in the previously determined SERCA crystal structures[1–5].

The five structures presented in the current papers[6–8] constitute a landmark in the history of investigations into P-type pumps. Pedersen *et al.*[8] (page 1111) report the E1 conformation of the proton pump AHA2 from the *Arabidopsis thaliana* plant, in complex with an analogue of ATP (Fig. 1a; the ATP analogue is chemically more stable than ATP itself, and so remains intact during crystallization). The pump generates the proton gradients (and so the electrical potentials) across the cell membrane that act as an energy source for plant and yeast cells. It belongs to a subfamily of P-type pumps — known as subfamily III — whose members have a relatively small binding domain for ATP. SERCA belongs to subfamily II, and so has a larger ATP-binding domain. But, aside from this expected difference, it can now be seen that the two pumps share the same arrangement of transmembrane helices and the same relative positions of their three cytoplasmic domains (Fig.1).

The second paper (Morth *et al.*[7], page 1043) reveals the E2 conformation of the sodium–potassium pump — a subfamily II protein — isolated from pig kidneys (Fig. 1b). The pump is in complex with an MgF_4^{2-} ion, which mimics a phosphate group, and occluded rubidium ions act as substitutes for the potassium ions that would be found *in vivo*. The sodium–potassium pump consists of three subunits, α, β and γ. The α-subunit is the largest, and is structurally similar to the entire SERCA protein. Morth and colleagues' structure[7] shows that, despite the presence of the β- and γ-subunits, the conformation of the α-subunit closely matches that of the analogous SERCA complex[4]. The two trapped rubidium ions are the first counter-transported ions large enough to be directly observed in the crystal structure of a P-type pump; they are in nearly the same positions as those thought to be adopted by the counter-transported H^+ ions in SERCA.

The structural differences between SERCA and the α-subunit of the sodium–potassium pump are remarkably small. They can be attributed to minor differences in the proteins' complement of amino acids, specific interactions of the α-subunit of the sodium–potassium pump with its other subunits, and the need for the sodium–potassium pump to coordinate two rubidium (or potassium) ions instead of two or three protons. Taken together with the previously obtained SERCA structures[1–5], the structure of AHA2 (ref. 8) and that of the sodium–potassium pump[7] show us that P-type pumps — or, at least, those in subfamilies II and III — share the same architecture regardless of the size, charge or number of ions that they transport, and that their differences are largely confined to the ion-binding pocket.

Olesen *et al.*[6] (page 1036) provide yet more bounty in the form of three new SERCA structures. The first of these shows SERCA in the E1P conformation, with two occluded calcium ions and an ADP-mimic bound to the protein. The phosphate group is firmly integrated into the pump protein, making this a true E1P state of SERCA.This is in contrast to previous structures[3,4] that were only 'E1P-like' because their phosphate mimics were more loosely attached.

Olesen and colleagues' other two structures[6] show E2 forms of SERCA. Previously obtained crystals of SERCA in the E2 state[2,4,5] required a pump inhibitor (such as thapsigargin) to be bound to the protein to stabilize the structure. But the new crystals[6] required no inhibitors, and so the structures are more likely to represent natural E2 conformations of SERCA. In fact, one of the new E2-SERCA structures (Fig. 1c) looks rather like an analogous E2-SERCA complex that was bound with thapsigargin[5]. It therefore seems that thapsigargin doesn't distort occluded E2-SERCA structures as had been feared.

But the crowning glory of this work[6] is a structure determined in the presence of a phosphate mimic (BeF_3^-), revealing a long-sought conformation of SERCA. Previously obtained biochemical data suggested that calcium ions bind weakly to SERCA pumps treated with BeF_3^-; such weak binding is consistent with the presence of an open ion pathway in the pumps[11]. This is finally confirmed by Olesen and colleagues' structure of the SERCA–BeF_3^-complex (Fig. 1d), in which the transmembrane helices are splayed apart, creating a funnel-shaped pathway. At its narrow end, this pathway leads to some of the ion-binding amino-acid

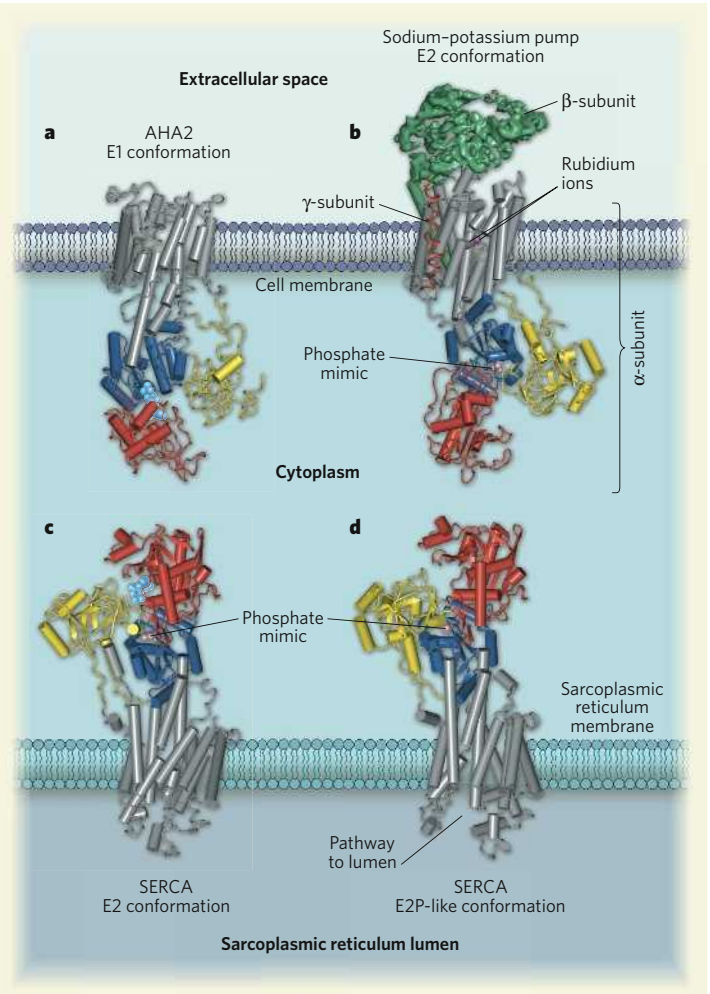


Figure 1 | Three P-type ion pumps have similar structures. P-type ion pumps transport ions across either cell membranes (**a**,**b**) or membranes of intracellular organelles such as the sarcoplasmic reticulum(**c**,**d**). **a**, Pedersen *et al.*[8] report the crystal structure of the AHA2 proton pump in the E1 conformation — ion-binding sites in the transmembrane region (grey rods) bind protons (positions not established) from the cytoplasm. **b**, Morth *et al.*[7] report the structure of the sodium–potassium pump in the E2 conformation. This comprises an α-subunit (which is structurally similar to the other pumps illustrated), a β-subunit (green) and a γ-subunit (single red helix). The pump is in complex with a phosphate mimic (MgF_4^{2-}), and two rubidium ions are trapped in the transmembrane region. **c**,**d**, Olesen *et al.*[6] report new structures of SERCA, a calcium ion pump. **c**, Here, SERCA is in the E2 conformation, in complex with AlF_4^- (another phosphate mimic). Two or three protons (positions not established) are trapped in the transmembrane region. **d**, This structure shows SERCA in complex with the BeF_3^- phosphate mimic. The conformation imitates the E2P state of the pump — a pathway is open to the lumen. In all cases the ATP-binding domains are red, phosphorylation domains are blue, key connecting regions to the transmembrane domain are yellow and ATP analogues (where bound) are pale blue. (Crystal structure images were prepared by A. Takeuchi.)

residues, which would thus be exposed to the lumen of the sarcoplasmic reticulum *in vivo*.

The five new structures[6–8] answer many long-standing questions about P-type ion pumps, but they also prompt further questions. The similarity between the analogous E2 forms of SERCA and of the α-subunit of the sodium–potassium pump suggests that SERCA structures will be useful models for other P-type pumps in subfamily II. But the overall similarity of the AHA2, SERCA and sodium–potassium-pump structures raises the question of how each protein selects only its preferred ions for transport. Higher-resolution structures of the pumps, in at least their two occluded conformations, are required to answer this question.

Issues specifically concerning the sodium–potassium pump also remain unresolved. In cells, this pump exports three sodium ions at a time from the cytoplasm, but then imports only two potassium ions, which are bound as seen in the new structure[7]. Other structures are sorely needed if we are to learn how the three sodium ions are handled. Furthermore, electrical signals generated by this pump during sodium-ion release suggest that the three sodium ions leave at different speeds, one after the other[12]. Does this mean that there is more than one escape route for these ions, implying that there might be more than one open-pathway conformation of the pump? SERCA also releases its two calcium ions sequentially[13], so might another open E2P conformation of this pump exist? Perhaps more fundamentally, the locations of the cytoplasmic ion pathways remain unidentified.

The remarkable structures reported today[6–8] will undoubtedly whet biologists' appetites. Would it seem greedy to ask for more ion-pump structures for Christmas? ■

David C. Gadsby is at the Laboratory of Cardiac/Membrane Physiology, The Rockefeller University, 1230 York Avenue, New York, New York 10065-6399, USA.
e-mail: gadsby@rockefeller.edu

1. Toyoshima, C., Nakasako, M., Nomura, H. & Ogawa, H. *Nature* **405,** 647–655 (2000).
2. Toyoshima, C. & Nomura, H. *Nature* **418,** 605–611 (2002).
3. Sørensen, T. L.-M., Møller, J. V. & Nissen, P. *Science* **304,** 1672–1675 (2004).
4. Toyoshima, C., Nomura, H. & Tsuda, T. *Nature* **432,** 361–368 (2004).
5. Olesen, C., Sørensen, T. L.-M., Nielsen, R. C., Møller, J. V. & Nissen, P. *Science* **306,** 2251–2255 (2004).
6. Olesen, C. *et al. Nature* **450,** 1036–1042 (2007).
7. Morth, J. P. *et al. Nature* **450,** 1043–1049 (2007).
8. Pedersen, B. P., Buch-Pedersen, M. J., Morth, J. P., Palmgren, M. G. & Nissen, P. *Nature* **450,** 1111–1114 (2007).
9. Post, R. L., Sen, A. K. & Rosenthal, A. S. *J. Biol. Chem.* **240,** 1437–1445 (1965).
10. Patlak, C. S. *Bull. Math. Biophys.* **19,** 209–235 (1957).
11. Danko, S., Yamasaki, K., Daiho, T., Suzuki, H. & Toyoshima, C. *FEBS Lett.* **505,** 129–135 (2001).
12. Holmgren, M. *et al. Nature* **403,** 898–901 (2000).
13. Forge, V., Mintz, E., Canet, D. & Guillain, F. *J. Biol. Chem.* **270,** 18271–18276 (1995).

MOLECULAR BIOLOGY

Genome under surveillance

Karen M. Arndt

Decoding the information stored in DNA requires an intricate balance between processes that turn gene expression on or off. A protein that influences the packaging of DNA regulates this balance genome-wide.

Organisms store instructions for their own existence in DNA. Specific proteins access and read the DNA sequence either to replicate it or to mediate gene expression. But this DNA-reading process is impeded by chromatin — tight packages of DNA and histone proteins that are essential for nuclear compartmentalization of the genome. Strategies for opening chromatin are therefore crucial for basic molecular processes such as gene transcription; this is not to say that restricting access to the genome is less important. On page 1031 of this issue, Whitehouse *et al.*[1] describe an elegant chromatin-based mechanism by which yeast (*Saccharomyces cerevisiae*) cells prevent inappropriate transcription.

Once considered static, chromatin is now viewed as a dynamic structure that regulates almost all aspects of DNA metabolism and genome inheritance. On a local scale, the positioning of nucleosomes (fundamental units of chromatin, comprising octamers of histone proteins wrapped by the DNA double strand) profoundly affects DNA-binding proteins' access to their target sequences. On a global scale, nucleosome positioning is non-random. For example, promoter sequences, which control transcription, are typically nucleosome-deficient, whereas coding regions tend to be nucleosome-rich[2,3]. Several strategies probably determine such consistent nucleosomal patterns. These include nucleotide sequences[4,5] and structural features of DNA[3] that disfavour or favour nucleosome formation, and active mechanisms such as competition between transcription factors and histones for binding to DNA, and enzyme-mediated chromatin remodelling[6].

Despite having been discovered almost two decades ago, the global effects of chromatin-remodelling factors are not fully understood. Although several studies had measured the transcriptional impact of mutating a chromatin-remodelling factor[6], it remained unknown whether these transcriptional effects are direct or indirect, or whether they are associated with changes in chromatin structure.

One chromatin-remodelling factor is the yeast Isw2 protein, which is a member of an evolutionarily conserved group of enzymes that alter chromatin structure using ATP as a source of energy. Different chromatin-remodelling factors alter chromatin structure in different ways, including nucleosome sliding, nucleosome assembly and disassembly, and

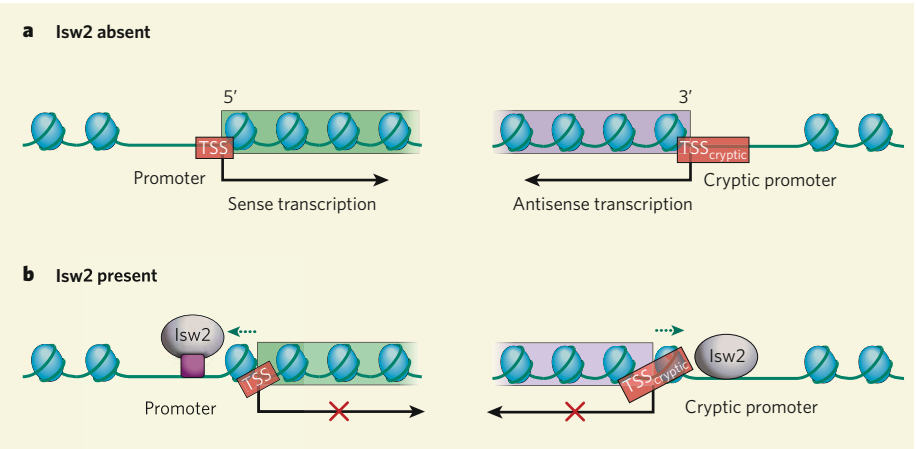


Figure 1 | The Isw2 protein represses transcription by altering nucleosome positions. **a**, Whitehouse *et al.*[1] show that, in yeast strains lacking Isw2 activity, nucleosomes are shifted towards coding regions at either the 5′ or 3′ ends of genes, and transcription is initiated either in the sense or antisense direction, respectively. TSS, transcription start site. **b**, Normally, specific DNA-binding proteins (magenta) recruit Isw2 to sequences within promoter regions, which are located within intergenic regions on the 5′ side of the coding regions. There, Isw2 slides nucleosomes towards intergenic regions, over sequences required for efficient initiation of transcription, such as the transcription start site. Isw2 is also recruited near the 3′ end of genes through an unknown mechanism. There, it also directs nucleosomes towards intergenic sequences, which may harbour cryptic signals for initiation of antisense transcription. So Isw2 functions as a transcriptional repressor.

histone exchange[6]. The enzymes that mediate these modifications are often members of large protein complexes that are recruited to DNA by site-specific binding proteins to facilitate or impede a biological process. In the case of yeast Isw2, its best-understood functions are sliding nucleosomes[6] — even onto energetically unfavourable DNA sequences[7] — and repressing transcription[8]. But, like similar proteins in multicellular eukaryotes, the functions of Isw2 probably extend to processes other than transcription[6].

Studying Isw2, Whitehouse *et al.*[1] provide the first global analysis of nucleosome positioning in a eukaryote lacking a chromatin-remodelling factor. The authors used genome-tiling arrays to map nucleosome positions in both normal and Isw2-deficient yeast strains. They then compared this analysis with genome-wide measurements of RNA levels in *ISW2* mutants and the localization of the Isw2 protein in the normal strain. The result is a comprehensive, high-resolution view of where on the yeast genome Isw2 localizes, where it directs changes in chromatin and where it alters transcription.

A fascinating picture emerges: the role of Isw2 in altering nucleosome position is global, with this remodelling factor regulating chromatin patterns in more than 1,000 genomic regions. Isw2 physically associates with many of these regions, which implies that it has direct effects on chromatin and transcription. Although removal of Isw2 might have altered nucleosome positioning at random locations in the genome, Whitehouse *et al.* observe a much more interesting pattern. They find a strong bias for Isw2 action and association at the 5′ and 3′ ends of genes. Strikingly, *ISW2* deletion caused nucleosomes that normally lie at the boundary of intergenic and coding regions to shift towards the coding regions. This shift was detected irrespectively of whether Isw2 was functioning at the 5′ or the 3′ end of a gene. Like a sliding-tile puzzle, the normal function of Isw2 is to slide nucleosomes towards intergenic regions, thus restricting the accessibility of these regions (Fig. 1).

At the 5′ end of genes, Isw2 slides nucleosomes towards the promoter, potentially obscuring the transcription start site, as well as the binding sites for regulatory factors; this would repress transcription. But what is the function of Isw2 at the 3′ end of genes? In strains lacking Isw2, the authors observed increased antisense transcription at the 3′ ends of several genes (that is, transcription in the opposite direction to the gene sequence). So Isw2 seems to play a central part in guarding the transcriptional integrity of the genome by preventing inappropriate initiation of transcription from within intergenic regions both in the sense and antisense directions.

Whitehouse and colleagues' work also raises many questions. For example, how is the function of Isw2 focused on the borders between intergenic and coding sequences? Although certain DNA-binding repressor proteins are known to recruit Isw2 to promoters[8], the large number of genes affected by the absence of this chromatin-remodelling factor indicates that other mechanisms exist. Does the histone protein H2A.Z, which selectively occupies the first nucleosome downstream of the transcription start site[9], have a part in Isw2 recruitment or activity? What mechanisms localize Isw2 to the 3′ ends of genes? Is the previously reported connection between Isw2 and termination of transcription[10] due to the effects of this factor on nucleosome positioning or on antisense transcription?

From a mechanistic standpoint, understanding how Isw2 generates a directional shift in nucleosomes, how its effects are limited to the first three nucleosomes of a gene, and how it blocks transcription will be exciting problems to address. It is also hoped that the insight gained by this comprehensive analysis[1] of one chromatin-remodelling factor will motivate similar studies on other such factors. Undoubtedly, more surprises are in store. ■

Karen M. Arndt is in the Department of Biological Sciences, 269 Crawford Hall, University of Pittsburgh, 4249 Fifth Avenue, Pittsburgh, Pennsylvania 15260, USA.
e-mail: arndt@pitt.edu

1. Whitehouse, I., Rando, O. J., Delrow, J. & Tsukiyama, T. *Nature* **450,** 1031–1035 (2007).
2. Yuan, G.-C. *et al. Science* **309,** 626–630 (2005).
3. Lee, W. *et al. Nature Genet.* **39,** 1235–1244 (2007).
4. Segal, E. *et al. Nature* **442,** 772–778 (2006).
5. Ioshikhes, I. P., Albert, I., Zanton, S. J. & Pugh, B. F. *Nature Genet.* **38,** 1210–1215 (2006).
6. Saha, A., Wittmeyer, J. & Cairns, B. R. *Nature Rev. Mol. Cell Biol.* **7,** 437–447 (2006).
7. Whitehouse, I. & Tsukiyama, T. *Nature Struct. Mol. Biol.* **13,** 633–640 (2006).
8. Goldmark, J. P., Fazzio, T. G., Estep, P. W., Church, G. M. & Tsukiyama, T. *Cell* **103,** 423–433 (2000).
9. Raisner, R. M. & Madhani, H. D. *Curr. Opin. Genet. Dev.* **16,** 119–124 (2006).
10. Alén, C. *et al. Mol. Cell* **10,** 1441–1452 (2002).

MATERIALS CHEMISTRY

Cool conditions for mobile ions

Michael A. Hayward and Matthew J. Rosseinsky

A complex iron oxide has been made that has an unusual crystal structure suggesting that the oxide ions are surprisingly mobile. This finding could pave the way to other metal-oxide materials with useful properties.

Transition-metal compounds provide a 'double whammy' of interest for researchers. Not only do they offer the fundamental scientific challenge of finding ways to control and enhance their properties, but they also have applications in such diverse areas as catalysis, magnetic information storage and battery technology. Reporting in this issue (page 1062), Tsujimoto *et al.*[1] describe how a previously unknown kind of mixed metal oxide can be made by side-stepping the thermodynamic limitations of traditional syntheses using a recently discovered method. This achievement opens the door to synthesis of many other complex metal oxides that have potentially useful properties.

The hallmark of transition metals is their ability to adopt a range of charged states, in which their outermost atomic orbitals (*d* orbitals) are only partly occupied by electrons. These electrons interact with ligands — neighbouring ions or molecules that bind to the metal — so that both the *d*-electron count and the number and arrangement of the ligands determine the electronic configuration of the metal. The electronic configuration in turn controls such properties as the colour or chemical reactivity of an isolated metal–ligand complex. When the transition-metal ions form part of a crystalline solid, ligands can be shared between metal atoms. This allows the *d* electrons on neighbouring metals to interact with each other, producing cooperative electronic properties[2] such as magnetic order or high-temperature superconductivity.

Ligands tend to have a limited repertoire of arrangements around metals, and these are determined by the number of *d* electrons available. In the search for materials with unusual properties, finding compounds with new arrangements of ligands around metals with a given number of *d* electrons is crucial. This is just what Tsujimoto *et al.*[1] have done in their discovery of $SrFeO_2$.

The authors' new compound is an example of a complex oxide — that is, an oxide that contains more than one type of metal cation, in this case strontium (Sr^{2+}) and iron (Fe^{2+}). Surprisingly, each Fe^{2+} ion is surrounded by four oxide ions (O^{2-}) arranged in a square, so that the whole complex is planar. This ligand geometry is expected to be highly unfavourable for Fe^{2+} ions in a complex oxide[3].

In the solid form of $SrFeO_2$, the iron oxide squares join together to form sheets like molecular patchwork quilts, in which oxide ligands are shared between Fe^{2+} ions (Fig. 1). The sheets, interleaved with layers of strontium ions to balance the overall charge, stack up to yield an 'infinite layer' structure. This ionic assembly is of great interest, because it is adopted by many copper oxide superconductors[4] — but it has never been observed for an iron oxide before. So why not?

To answer this question, one needs to

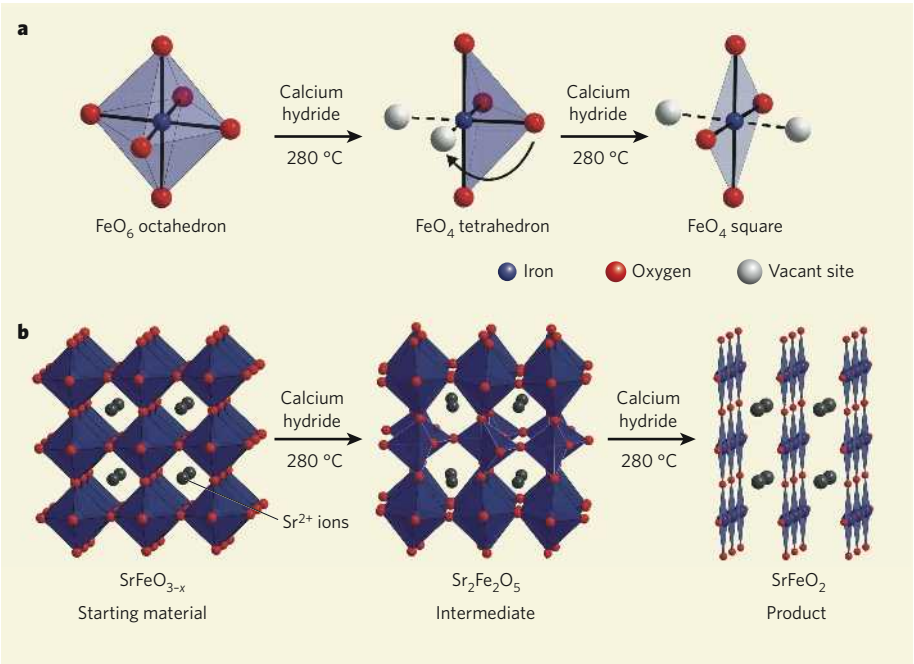


Figure 1 | Ion movement in a complex metal oxide. Tsujimoto *et al.*[1] report the preparation of $SrFeO_2$ (Sr is strontium, Fe is iron), a complex metal oxide with an unusual arrangement of ions. **a**, The starting material ($SrFeO_{3-x}$, where x is about 0.125) contains FeO_6 octahedra. On heating with calcium hydride, some of the oxygen ions (known as oxide ions) are lost, so that intermediate FeO_4 tetrahedra form. One of the remaining oxide ions then moves to a vacant site left behind by an ion (indicated by the arrow), forming an FeO_4 square. **b**, The crystal lattice of the starting material is an array of FeO_6 octahedra, in which each oxygen is shared between two iron atoms. Strontium ions (Sr^{2+}) fit in between the rows of octahedra. In the first step of the reaction, loss of some of the oxide ions leads to an intermediate compound, $Sr_2Fe_2O_5$, consisting of alternating rows of FeO_6 octahedra and FeO_4 tetrahedra. In the second step, more oxide ions are lost and some of the remaining oxide ions change position. The $SrFeO_2$ product thus forms as sheets of FeO_4 squares, interleaved with strontium ions. The iron and strontium ions retain their positions throughout the process.

consider the differences between solid-state reactions and those that occur in solution. Reactions in solution can be performed at low temperatures, because molecular diffusion occurs easily and the reacting molecules don't require much energy to mix together. Under these conditions, if a molecule can take part in several reactions, only the one with the lowest energy barrier to activation tends to occur. The product of such a 'kinetically controlled' reaction is the one that forms fastest, and is not necessarily the one that is most stable. It is therefore possible to control reactions in solution so that they occur only at specific parts of a molecule. By performing stepwise transformations on individual chemical groups, a product can be prepared that has a controlled composition and structure that are clearly related to those of the starting compound.

But nearly all complex metal oxides are prepared at high temperatures (typically greater than 1,000 °C). This is because no solvent is used to aid diffusion, yet the reacting ions must travel large distances (of the order of micrometres) to form the products. At these temperatures, enough energy is available to allow the occurrence of reactions that have high energy barriers. Given a choice of reaction pathways, the most favourable one is that which yields the most thermodynamically stable configuration of atoms. In such thermodynamically controlled systems, the product generally does not conserve any of the structural features of the reactants (unlike reactions in solution), and it is therefore much more difficult to direct the course of the reaction[5].

Tsujimoto *et al.*[1] overcome these limitations in their preparation of $SrFeO_2$. Their starting material is a complex metal oxide that contains Fe^{4+} ions ($SrFeO_{3-x}$, where x is about 0.125). The authors form their unusual product by removing an oxide ion from the starting material, a process that is coupled to a redox reaction in which Fe^{4+} ions are converted into Fe^{2+} ions.

The overall process uses a recently discovered reagent (calcium hydride) for the kinetically controlled removal of oxygen from oxides[6], and it occurs at the remarkably low temperature of 280 °C. These conditions provide insufficient thermal energy to rearrange the structure of the starting material completely — only the relatively mobile oxide ions can change position (Fig. 1). The strontium and iron ions in the product retain the positions they held in the starting material. The most thermodynamically stable products — iron metal and strontium(II) oxide — do not form, because the required long-range diffusion for the process is too slow at this temperature. The less stable $SrFeO_2$ forms instead, because this is a faster reaction.

Nevertheless, the reaction pathway that leads to $SrFeO_2$ (Fig. 1) is unexpected. An intermediate ($Sr_2Fe_2O_5$) is formed first, as oxide ions are removed from the starting material. This intermediate consists of alternating sheets of FeO_4 tetrahedra and FeO_6 octahedra. Conversion of the tetrahedra into the square planes of the final product requires that the oxide sites vacated in the formation of the intermediate be refilled with other oxide ligands. This is a crucial observation, because it demonstrates that all the oxide ions are mobile, not just those being removed from the system.

The discovery that oxide ions can be mobile at relatively low temperatures, albeit in the presence of a strong chemical driving force, opens up a host of synthetic possibilities — for example, the strong magnetic interactions seen in $SrFeO_2$ could be modified in a controlled way by making complex oxides of different transition metals. But the practical applications are just as exciting — high oxide-ion mobility is required for several emerging technologies, most notably solid-oxide fuel cells[7]. So although Tsujimoto and colleagues' discovery[1] may occur only at an atomic level, its ramifications could extend far more widely. ■

Michael A. Hayward is in the Inorganic Chemistry Laboratory, Department of Chemistry, University of Oxford, South Parks Road, Oxford OX1 3QR, UK.
Matthew J. Rosseinsky is in the Department of Chemistry, University of Liverpool, Liverpool L69 7ZD, UK.
e-mails: michael.hayward@chem.ox.ac.uk; m.j.rosseinsky@liverpool.ac.uk

1. Tsujimoto, Y. *et al. Nature* **450,** 1062–1065 (2007).
2. Blundell, S. *Magnetism in Condensed Matter* (Oxford Univ. Press, 2001).
3. Rao, C. N. R. & Raveau, B. *Transition Metal Oxides: Structure, Properties, and Synthesis of Ceramic Oxides* 2nd edn (Wiley-VCH, Weinheim, 1998).
4. Cava, R. J. *J. Am. Ceramic Soc.* **83,** 5–28 (2000).
5. Stein, A., Keller, S. W. & Mallouk, T. E. *Science* **259,** 1558–1564 (1993).
6. Hayward, M. A. *et al. Science* **295,** 1882–1884 (2002).
7. Atkinson, A. *et al. Nature Mater.* **3,** 17–27 (2004).

Corrections

- The News & Views article "Venus: Express dispatches" by Andrew P. Ingersoll (*Nature* **450,** 617–618; 2007) contained the erroneous statement that Venus's equator is warmer than the poles at altitudes above 65 km. It is colder.

- There was an incorrect reference citation in the article "Microscopy: Elementary resolution" by Christian Colliex (*Nature* **450,** 622–623; 2007). In the statement "The first experimental maps are now demonstrating the importance of refining descriptions of electron–matter interactions[2]," the correct citation is not reference 2 but reference 12 (M. Bosman *et al. Phys. Rev. Lett.* **99,** 086102; 2007).

- In the article "Astronomy: Sloan at five" by Robert C. Kennicutt Jr (*Nature* **450,** 488–489; 2007), we should clarify that the Sloan Digital Sky Survey was used only to select candidate stars for the spectroscopic observations that led to the discovery cited in reference 9 (W. R. Brown *et al. Astrophys. J.* **622,** L33–L36; 2005).

OBITUARY

Gene H. Golub (1932–2007)

Mathematician and godfather of numerical analysis.

A century ago, matrices and the techniques for their manipulation — linear algebra — were a backwater of mathematics. Today, they are the foundation not just of the mathematical field of numerical analysis, but also of computational science and engineering, and have become indispensable for anyone who wants to get numerical results from a computer. The pre-eminent figure in matrix computations over the past 50 years, Gene Golub, died on 16 November.

Golub was born in Chicago on 29 February 1932, to Jewish parents from Latvia and the Ukraine. His childhood was not affluent, but he was a good student. After two years at a junior college, he transferred to the University of Illinois at Urbana-Champaign, achieving his doctorate there in 1959. At the time, Illinois, with the first of its 'ILLIAC' supercomputers, was a great centre of computing, and Golub showed his affection for his Alma Mater by endowing a chair there 50 years later. Rumour has it that the funds for the gift came from Google stock acquired in exchange for some advice on linear algebra. Google's PageRank search technology starts from a matrix computation — an eigenvalue problem with dimensions in the billions. Hardly surprising, Golub would have said: everything is linear algebra.

He came to believe that in his twenties, as he realized that new methods of orthogonal-matrix factorization introduced by Wallace Givens and Alston Householder offered the right mathematical recipe for solving all kinds of problems. In particular, Golub focused on the idea known as singular value decomposition, SVD, which systematically isolates the dominant components of a linear process. Together with William Kahan and Christian Reinsch, he invented the now-standard SVD algorithms, and showed scientists, engineers and statisticians how these algorithms could be used in areas such as the least-squares method to find the best fit to a curve; in optimization problems and control theory; and for the determination of crucial matrix parameters such as their norms, ranks and condition numbers. In later years he drove a car with the licence plate 'PROF SVD'.

Golub found his way to Stanford University in 1962, eventually becoming the senior professor in its formidable computer science department. In 45 productive years there, he advanced matrix computations in areas as diverse as geodesy, data mining and quantum chromodynamics. The dimension of what was considered a 'big' matrix grew from 100 to 1,000,000 in the same period, and Golub was among the first to develop the iterative algorithms that make problems involving such huge matrices tractable.

As the new methods came in, older ideas such as gaussian elimination (essentially, the way one is taught to solve a system of simultaneous equations in school, by eliminating the variables one by one) became a smaller part of a new and greater enterprise. Along with the new algorithms came a new world of software for solving mathematical problems, such as EISPACK, LAPACK and MATLAB. Golub's book *Matrix Computations*, co-authored with Charles Van Loan of Cornell University, became a bestseller and the definitive textbook of the field. Honours flowed in, including membership of the US National Academies.

As a servant of the wider scientific community, Golub did as much as anybody to make the Society for Industrial and Applied Mathematics (SIAM) the organization it is today. He served it in various capacities, among them as president (1985–87). He also founded and edited two of the society's journals, the *SIAM Journal on Scientific and Statistical Computing* and *SIAM Journal on Matrix Analysis and Applications*. It was his proposal that led to the quadrennial International Congresses on Industrial and Applied Mathematics.

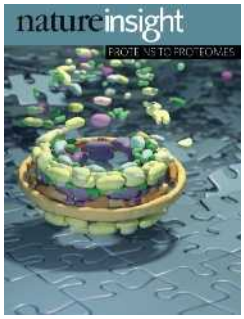
But this impressive list of achievements misses the truly extraordinary aspect of this complex man: the scale of his devotion to people. Golub was a bachelor for most of his life, and his colleagues were his family. No family ever had a more loving, attentive or exasperating father. As he liked to say, "Every numerical analyst has a second home at Stanford". Countless colleagues enjoyed a glass of wine at his home there, and hundreds of them stayed over for a night or even a month at his invitation. How did he remember all our birthdays and reading tastes and children's names?

Golub could not spend a day without other people. He would eat dinner with them, talk matrices with them, organize conferences with them, write papers and books with them, argue academic politics with them — an endless dance of interactions, plans and projects. Anywhere in the world, a numerical analyst knows who is meant by 'Gene'. About 250 of them were his co-authors. They knew that it would fall to them to do most of the writing; but Golub saw the connections, knew the literature, and made the paper happen.

He seemed almost to have invented e-mail. As early as 1981, his office computer was set up to beep the moment a message arrived. His personal address list evolved into the worldwide database of numerical analysts, and his notes to friends became the Numerical Analysis Digest. This newsletter, one of the first e-bulletins, is now sent to some 8,000 recipients weekly. He could not sit still. As he left us in Oxford last September after an extended sabbatical visit, having spent much of the preceding months talking with the graduate students in the common room — to which he had donated $1,000 for a biscuit fund — he mentioned that he had three trips to China planned for the upcoming year.

Gene Golub was restless and never entirely happy. He was a demanding friend; behind his back, we all had Gene stories to tell. It was a huge back: Gene was big, dominating any room he was in, and grew more impressive and imposing with the years. Graduate students around the world admired and loved him, and he bought them all dinner when he got the chance. His unexpected death, in Stanford in between speaking at a conference in Hong Kong and flying to Zurich for his eleventh honorary degree, has left the world of numerical analysis orphaned and reverberating.

Lloyd N. Trefethen
Lloyd N. Trefethen is in the Oxford University Computing Laboratory, Oxford OX1 3QD, UK.
e-mail: lnt@comlab.ox.ac.uk



Cover illustration
The molecular architecture of the nuclear pore complex. (Courtesy of S. Parker and C. Silva, Scientific Computing and Imaging Institute, and R. K. Morley, RayScale).

Editor, *Nature*
Philip Campbell
Insights Publisher
Sarah Greaves
Publishing Assistant
Claudia Banks
Insights Editor
Ritu Dhand
Production Editor
Davina Dadley-Moore
Senior Art Editor
Martin Harrison
Art Editor
Nik Spencer
Sponsorship
Emma Green
Production
Jocelyn Hilton
Marketing
Katy Dunningham
Elena Woodstock
Editorial Assistant
Alison McGill

PROTEINS TO PROTEOMES

How are innumerable protein functions integrated so that a living cell interacts coherently with its environment? This question is central to an emerging science of biological information processing — systems biology.

For nearly two centuries, 'physiological chemists' had to tackle the opposite problem: making sense of the murky 'protoplasm' of cells by dividing it into separate components. From the discovery in 1840 of haemoglobin crystals in earthworm blood spread on glass plates, to the development of the ultracentrifuge in the 1920s, to contemporary genomics research, biochemists have strived to study life's parts in isolation. Now that part lists are almost complete, the pieces must be brought back together.

High-throughput proteomics has enabled large protein-interaction networks to be drafted, and the statistical analysis of these forms a busy branch of systems biology. But rather than just a 'dot' connected in a network, each protein is a complex and dynamic three-dimensional object with sophisticated chemical properties. And proteins can assemble into large, asymmetrical or transient molecular machines. Accurate descriptions of these assemblies, in turn, must reflect the cellular context in which they usually operate, a factor that also conditioned their evolution.

Cutting-edge protein science therefore remains central to any meaningful modelling of biological systems. This Insight covers some of the most vibrant areas of research into the 'protein world', taking a journey from single protein dynamics to functional proteomics and drug discovery, through some of the latest technological developments in structural, cellular, evolutionary and computational biology.

We are pleased to acknowledge the financial support of Pfizer in producing this Insight. As always, *Nature* carries sole responsibility for all editorial content and peer review.

Tanguy Chouard and Joshua Finkelstein, Senior Editors

REVIEWS

nature insight

NATURE|Vol 450|13 December 2007|doi:10.1038/nature06522

Dynamic personalities of proteins

Katherine Henzler-Wildman[1] & Dorothee Kern[1]

Because proteins are central to cellular function, researchers have sought to uncover the secrets of how these complex macromolecules execute such a fascinating variety of functions. Although static structures are known for many proteins, the functions of proteins are governed ultimately by their dynamic character (or 'personality'). The dream is to 'watch' proteins in action in real time at atomic resolution. This requires addition of a fourth dimension, time, to structural biology so that the positions in space and time of all atoms in a protein can be described in detail.

Life is marked by change over time, and biologists explore this phenomenon by watching, for example, *Caenorhabditis elegans* developing from embryos into adults, mice running in a cage, and nerve cells firing. In search of how and why, biology arrived at the molecular level. Understanding protein function on an atomic level has been revolutionized by high-resolution X-ray crystallography, resulting in a surge in studies of structure–function relationships. The detail in these colourful structures flooding the covers of modern journals can be deceptive, suggesting that one unique structure, the 'folded state', is the final answer. Ironically, the dynamic nature of biology seems to have been forgotten at this microscopic level.

Physicists, however, will object to a static picture: they see proteins as soft materials that sample a large ensemble of conformations around the average structure as a result of thermal energy. A complete description of proteins requires a multidimensional energy landscape that defines the relative probabilities of the conformational states (thermodynamics) and the energy barriers between them (kinetics). In biology, this concept has recently gained traction, leading to an extension of the structure–function paradigm to include dynamics. To understand proteins in action, the fourth dimension, time, must be added to the snapshots of proteins frozen in crystal structures. A major obstacle is that it is not possible to watch experimentally individual atoms moving within a protein. Instead, sophisticated biophysical methods are needed to measure the physical properties from which the dynamics can be inferred.

In this review, we discuss how protein function is rooted in the energy landscape. The basic concepts and the biophysical methods are illustrated by several examples. To avoid past semantic confusion about the term protein dynamics, we define it as any time-dependent change in atomic coordinates. Protein dynamics thus includes both equilibrium fluctuations and non-equilibrium effects. The fluctuations observed at equilibrium seem to govern biological function in processes both near and far from equilibrium; therefore, we focus on these motions. Non-equilibrium effects are also called dynamical effects[1] (the source of confusion[2]), and they have a minimal effect on the overall rates of biological processes[3,4]. Biological motors that convert chemical energy to mechanical energy[5,6] are not discussed here.

The energy landscape

Although the idea of an energy landscape might be most familiar in the context of protein folding (for example, the folding funnel hypothesis)[7–9], this concept had already been applied to folded proteins more than 30 years ago by Frauenfelder and co-workers[10]. Using flash photolysis, they measured the kinetics of carbon monoxide and oxygen rebinding to myoglobin as a function of temperature and ligand concentration[10]. Based on the observation of multiple energy barriers and non-exponential kinetics below a temperature of 230 K, an energy-landscape model was developed[11]. Frauenfelder and colleagues insightfully connected this energy-landscape concept to myoglobin function and characterized the features of the landscape, including the heights of the barriers between energy wells and the existence of multiple conformational substates[12]. Subsequent studies on myoglobin led to the idea that substates are in thermal equilibrium and that both solvent[13–15] and ligands influence the landscape (Fig. 1a). At the glass transition temperature[10,12], an increase in anharmonic dynamics occurs in proteins, and this is interpreted as the protein no longer being trapped in a single energy well. This transition has recently been attributed to a solvent relaxation effect in the hydration shell of proteins[16]. Since these early studies, many more details of protein energy landscapes have been characterized as a result of advances in experimental and computational techniques (described later).

We divide our discussion based on the timescale of the dynamic processes (Fig. 1). It should be noted, however, that protein dynamics are characterized not only by the timescale of the fluctuations (a kinetic component) but also by the amplitude and the directionality of the fluctuations (a structural component). Consequently, the energy landscape representing a protein, which has many atoms, is highly multidimensional. It is also important to keep in mind that a particular energy landscape is tied to an individual set of temperature, pressure and solvent conditions. Manipulating these conditions is one of the most common ways to change the relative populations of the states and the kinetics of conversion between them. Logically, the energy landscape of a protein is inclusive of all the states sampled by the protein–solvent system, including the unfolded subspace. The process of protein folding, however, has been discussed thoroughly elsewhere (see refs 7–9, 17 and 18 for reviews) and is not covered here.

Slow timescales

Dynamics on a 'slow' timescale (tier-0 dynamics) define fluctuations between kinetically distinct states that are separated by energy barriers of several kT (the product of the Boltzmann constant and the absolute temperature), corresponding to timescales of microseconds and slower at physiological temperature. Typically, these are larger-amplitude collective motions between relatively small numbers of states. The protein is not static within one of these tier-0 states; instead, it fluctuates around the average structure on a faster timescale, exploring a large ensemble of closely related structures (see the section Fast timescales).

[1]Department of Biochemistry, Howard Hughes Medical Institute, Brandeis University, Waltham, Massachusetts 02454, USA.

Transitions between tier-0 states are rare, however, owing to the low probability of the conformation that allows transition. Dynamics on this timescale have received much attention recently, because many biological processes — including enzyme catalysis, signal transduction and protein–protein interactions — occur on this timescale. Owing to the relatively long lifetimes of each state, these individual states can either be observed directly or be trapped experimentally. Moreover, the kinetics of interconversion of these states can also be detected. In this section, we discuss what has been learned about dynamics on slow timescales from experimental atomic-resolution methods, experimental low-resolution and local-site methods, and computational methods.

Experimental atomic-resolution methods

Ideally, researchers would like to determine both the structures of the tier-0 substates and their rates of interconversion. X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, cryo-electron microscopy and small-angle X-ray scattering provide atomic-resolution or near-atomic-resolution snapshots of tier-0 substates. For high-resolution X-ray crystallography, a homogeneous crystal is needed. Consequently, substates need to be trapped through biochemical 'tricks', or the reaction needs to be synchronized across the entire crystal[19]. These ideas are nicely illustrated by the crystallographic characterization of intermediates in the cytochrome P450 enzymatic cycle[20].

The requirement for a homogeneous crystal is relieved when using cryo-electron microscopy and small-angle X-ray scattering, making it possible to determine the structural ensemble directly, in the experimental conditions, although with lower resolution. However, these methods cannot characterize the timescales of interconversion. Usually, this structural information is linked to kinetic data obtained from low-resolution spectroscopic methods (discussed in the next subsection, Experimental low-resolution and local-site methods). In specialized circumstances, both structures and kinetics can be determined simultaneously by using Laue X-ray diffraction[19]. In addition, hydrogen–deuterium exchange, analysed by either mass spectrometry or NMR spectroscopy, provides a particularly powerful way to detect global or local unfolding on timescales of milliseconds and longer[21,22].

The clear advantage of NMR methods is that they deliver the timescale of transitions, together with atomic resolution. Dynamic information is extracted from relaxation of nuclei after excitation, using a variety of NMR experiments to span dynamics on timescales from picoseconds to seconds and to assess several types of nucleus (^{1}H, ^{2}H, ^{13}C and ^{15}N) site specifically[23–25]. Importantly, the dynamics can be followed in solution in steady-state conditions[26]. This is in contrast to most other spectroscopic methods, which require perturbation to measure kinetics. NMR experiments have traditionally been limited to small, soluble proteins. However, modern spectrometer technology (such as high magnetic fields and cryoprobes) and new NMR pulse sequences have pushed the size limit upward, making it possible to study proteins of up to 100 kDa and even up to the size of the ribosome, depending on the system and question of interest[27–33].

The NMR timescale for conformational exchange is defined by its rate (k_{ex}, the sum of the forward and reverse rates) relative to the chemical-shift timescale ($\Delta\omega$, the difference in chemical shift of the interconverting species). Interconversion is slow on the NMR timescale when $k_{ex} < \Delta\omega$, fast when $k_{ex} > \Delta\omega$, and intermediate when $k_{ex} \approx \Delta\omega$ (ref. 25). For a slow exchange rate, the substates are observed as distinct peaks in the spectrum, allowing direct structural characterization. The relative populations of the substates (p_A and p_B) are obtained from the relative peak integrals, and exchange rates from one-tenth of a second to tens of seconds can be measured by nuclear Överhauser enhancement spectroscopy (NOESY) and ZZ-exchange spectroscopy[24]. By contrast, at intermediate and fast exchange rates, a single population-averaged signal is obtained. Microsecond-to-millisecond dynamics cause additional line broadening of this signal by an amount, R_{ex}, that contributes to the measured overall transverse relaxation rate (R_{2eff}). Specialized relaxation dispersion experiments[24,25,34] have been developed, allowing

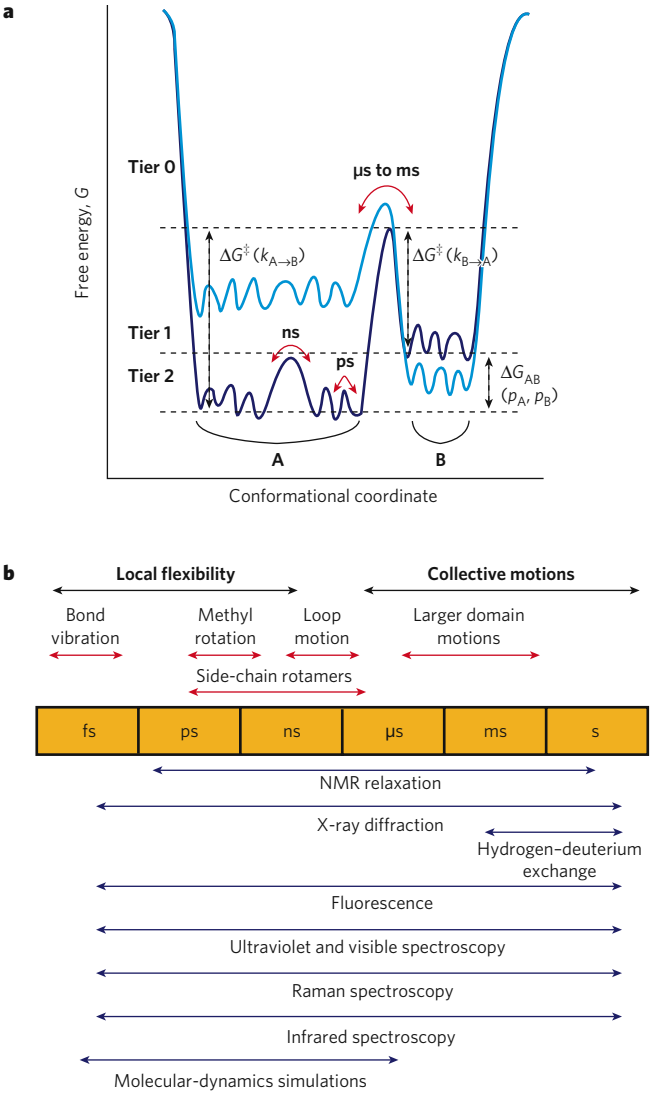


Figure 1 | The energy landscape defines the amplitude and timescale of protein motions. a, One-dimensional cross-section through the high-dimensional energy landscape of a protein showing the hierarchy of protein dynamics and the energy barriers. Each tier is classified following the description introduced by Frauenfelder and co-workers[93]. A state is defined as a minimum in the energy surface, whereas a transition state is the maximum between the wells. The populations of the tier-0 states A and B (p_A, p_B) are defined as Boltzmann distributions based on their difference in free energy (ΔG_{AB}). The barrier between these states ($\Delta G^{\ddagger}$) determines the rate of interconversion (k). Lower tiers describe faster fluctuations between a large number of closely related substates within each tier-0 state. A change in the system will alter the energy landscape (from dark blue to light blue, or vice versa). For example, ligand binding, protein mutation and changes in external conditions shift the equilibrium between states. **b**, Timescale of dynamic processes in proteins and the experimental methods that can detect fluctuations on each timescale.

determination of k_{ex} (kinetics), p_A and p_B (thermodynamics) and $\Delta\omega$ (structure) from the dependence of R_{ex} on an applied effective magnetic field (ν_{CPMG}) (Fig. 2a).

Using these dispersion experiments, the protein dynamics in an enzyme during catalysis have been measured, for cyclophilin A (CYPA)[35,36] (Fig. 2). CYPA catalyses the reversible *cis–trans* isomerization of prolyl peptide bonds. It was originally identified as the target of the immunosuppressive drug cyclosporin A[37,38]. Since then, peptidylprolyl isomerases have emerged as important regulators of various biological processes. For such a reversible enzyme, catalysis can be maintained indefinitely in the sample tube by simply adding the substrate(s)[26]. Quantitative analysis of the NMR dispersion experiments on CYPA[34,39,40]

revealed a global conformational exchange process that coincides with the chemical step of peptidylprolyl isomerization of the substrate on the enzyme[36] (Fig. 2a, b). The dynamics of individual microscopic steps of the catalytic cycle were dissected (that is, binding and dissociation, and isomerization)[35], and the collective nature of motions in a large dynamic network was experimentally characterized by studying proteins with various mutations[36] (Fig. 2d). Strikingly, characteristic motions detected during catalysis are already present in the free enzyme with frequencies similar to the turnover numbers (the number of molecules of substrate converted to product by one enzyme site per second)[36] (Fig. 2c). Therefore, the dynamics are an intrinsic property of the enzyme that is 'harvested' for catalytic turnover[36]. We propose that this fundamental finding that free CYPA 'pre-samples' the conformational substates observed during catalysis might be a general paradigm for enzymes.

Experimental low-resolution and local-site methods

In the era of atomic-resolution methods, the classical biophysical techniques of fluorescence, circular dichroism, absorbance, infrared spectroscopy, Raman spectroscopy and electron paramagnetic resonance have been treated as second-class citizens. However, these time-honoured methods are now having a renaissance, owing to an appreciation of their power to provide kinetic information that is complementary to higher-resolution methods. These lower-resolution methods access a large range of timescales (Fig. 1b) with high precision, while providing information for one or a few sites or an average over the entire system.

Here, we focus on one new and exciting area in this category, the application of fluorescence at the single-molecule level[41–44]. This technique brings to life a dream that biochemists have had for many years — watching a single protein molecule functioning in real time. The observed lifetimes of these states typically follow statistical exponential distributions, which are manifested at the macroscopic level in the familiar exponential kinetics that have been measured for ensembles. The strength of single-molecule methods is their ability to detect molecular heterogeneity, transient intermediates, rare events and the sequence of events, all of which might be hidden in population-averaged measurements. Fluorescence methods can detect single molecules because of the high sensitivity afforded by optimized optics and fluorescent dyes, combined with efficient detectors and detection geometries. In addition, fluorescence resonance energy transfer (FRET) can serve as a 'spectroscopic ruler'[45], allowing characterization of distance over time when experiments are carried out in a time-resolved manner. One of the limitations of single-molecule FRET is that only a single distance change is measured. When higher-resolution structural information is available, however, these distance changes can be interpreted in terms of possible corresponding conformational changes.

The power of single-molecule FRET to unravel the detailed molecular mechanism of an important multisubunit enzyme was elegantly demonstrated by Diez and colleagues for F_0F_1-ATP synthase[46] (Fig. 3). This membrane-bound enzyme converts the electrochemical energy of a transmembrane proton gradient into chemical energy in the form of ATP. Using the intact protein complex in vesicles (liposomes), the stepwise rotation of the γ-subunit driven by the proton gradient was followed in real time by single-molecule FRET (Fig. 3). The slow diffusion of the particles, due to the size of the liposomes, allowed measurements over hundreds of milliseconds. Strikingly, the experiment captured several rotational steps between three distinct FRET levels in a specific order. The order of switching was reversed during ATP synthesis, relative to ATP hydrolysis. This heroic single-molecule FRET study provided insights into the mechanism of catalysis that could only be obtained by single-molecule experiments, owing to the consecutive and progressive nature of the dynamics involved. An important feature exposed by such single-molecule experiments is that the actual transition between the substates is fast (faster than the time resolution), whereas the observed 'slowness' of switching arises from the low probability of transitions (Fig. 3c).

Computational methods

Computation has the unbeatable edge in that it can describe protein dynamics completely: the precise position of each atom at any instant in time for a single protein molecule can be followed, along with the corresponding energies, provided that at least one high-resolution structure is known as a starting point. Although conformational substates (located in energy wells) and their rates of interconversion can be detected experimentally (as described earlier), an atomic-resolution structural description of the 'climb from one valley to another' (the transition pathway) is out of experimental reach, owing to the extremely low probability and short lifetime of the high-energy conformers. Computational methods would be able to overcome these limitations if a perfect description of the protein–solvent system could be provided by

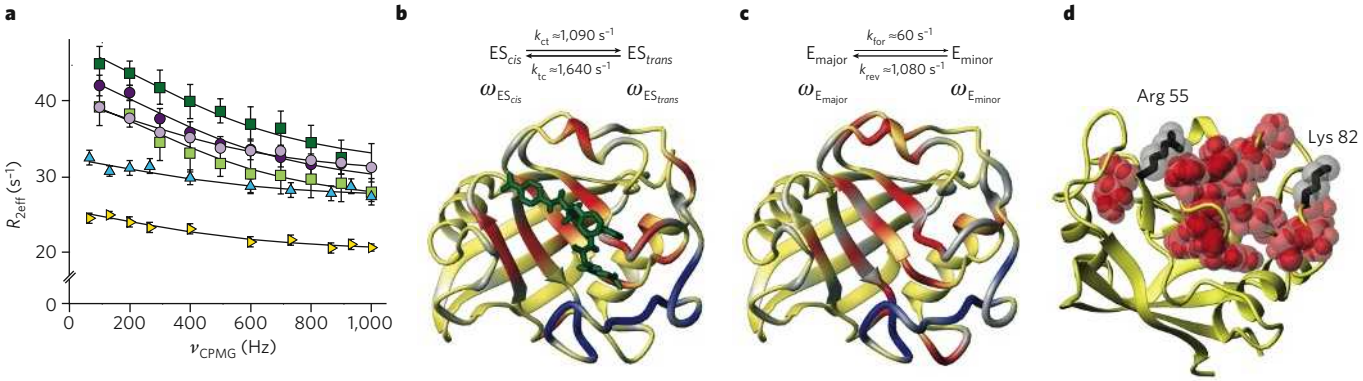


Figure 2 | Microsecond-to-millisecond protein dynamics are necessary for catalysis and are an intrinsic property of CYPA as shown by NMR relaxation dispersion experiments. a, Global fit of NMR relaxation data for representative ^{15}N backbone amides and ^{13}C methyl groups in CYPA (denoted by different shapes) during catalysis of the peptide *N*-succinyl-Ala-Phe-Pro-Phe-*p*-nitroanilide is shown. R_{2eff} is the overall transverse relaxation rate, and ν_{CPMG} is the applied effective magnetic field. **b**, During catalysis of the peptide (green) by CYPA, residues undergoing conformational exchange at the rates shown in the reaction scheme are plotted on the structure (red). Residues in one loop (blue) undergo exchange at a faster rate. Residues for which there are no data are shown in grey. ES denotes enzyme with substrate, and ω denotes chemical shift. The rate constant for *cis* to *trans* isomerization is denoted k_{ct}; and for *trans* to *cis* isomerization, k_{tc}. **c**, Analysis of the free enzyme, E, reveals a striking correspondence in the residues undergoing exchange on a similar timescale. Moreover, agreement between the chemical-shift differences of the exchanging species (between $\omega_{ES_{cis}} - \omega_{ES_{trans}}$ and $\omega_{E_{major}} - \omega_{E_{minor}}$) implies that exchange occurs between the same two states in each case and that the substrate merely shifts a pre-existing equilibrium. The rate constants for the forward and reverse reaction are denoted k_{for} and k_{rev}, respectively. **d**, Residues that build a common dynamic network (displayed as van der Waals radii, red) were identified by measuring chemical-shift changes between the wild-type protein and mutant proteins in which either Arg 55 or Lys 82 was mutated to alanine (residues shown in black). These chemical-shift changes are caused simply by shifting the pre-existing equilibrium. (Figure reproduced, with permission, from ref. 36.)

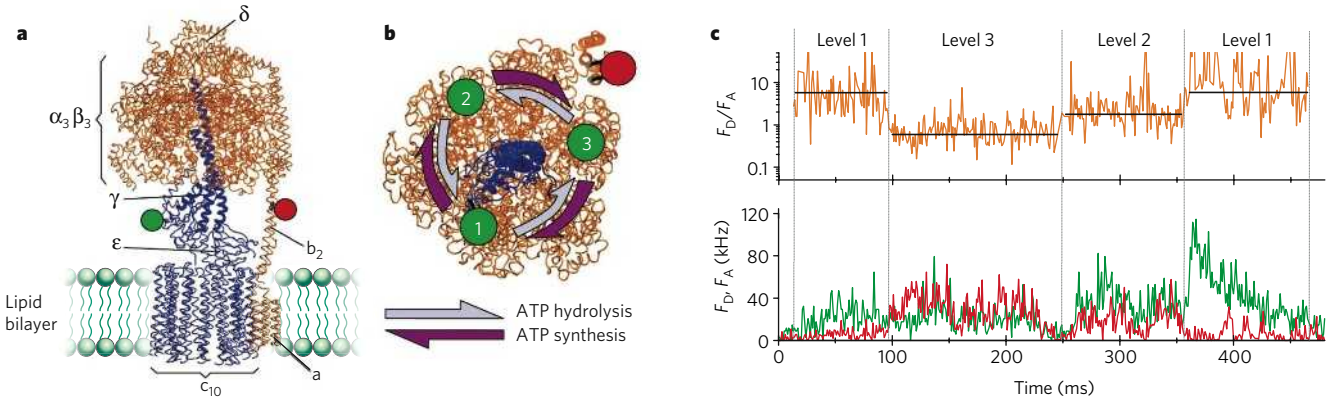


Figure 3 | Single-molecule FRET reveals ordered, stepwise rotation of F_0F_1-ATP synthase on the millisecond timescale during ATP hydrolysis and synthesis. a, Model of F_0F_1-ATP synthase embedded in a lipid bilayer. The rotor subunits are shown in blue, and the stator subunits are shown in orange. The FRET donor (green) is bound to the γ-subunit, and the FRET acceptor (red) is bound to the b-subunits. **b**, Cross-section at the level of the fluorophore, as viewed from the membrane. The change in position of the donor is shown relative to the acceptor on rotation of the rotor subunits (blue) in 120° steps. **c**, Single-molecule time traces of a single F_0F_1-ATP synthase molecule in liposomes during ATP hydrolysis. The fluorescence intensities (lower panel) of the donor (F_D, green) and the acceptor (F_A, red), as well as the corrected intensity ratio (F_D/F_A, upper panel, orange), uncover stepping between three states, with unique donor–acceptor distances in the order 1→3→2→1 (which correspond to the numbers in part **b**). Data were collected over 1 ms intervals. (Figure reproduced, with permission, from ref. 46.)

the force field (that is, parameter sets describing the potential energy of all atoms). Impressive progress has been made in the development of these force fields since their original conception[47,48], and they are used in molecular-dynamics simulations[47,49] (see the section Fast timescales).

Unfortunately, protein dynamics on the microsecond-to-millisecond timescale is currently out of reach for conventional molecular-dynamics simulations. To overcome this restriction, a large variety of approaches that simplify force fields have been developed, including normal mode analysis[50,51], gaussian network models[52], FIRST (floppy inclusion and rigid substructure topography)[53], FRODA (framework rigidity optimized dynamic algorithm)[54] and Gō models[49]. Alternatively, the dynamic process is accelerated by external force to access this timescale (used in methods such as targeted, steered and accelerated molecular-dynamics simulations[47,55–57]), or prior knowledge about features of the reaction coordinate (umbrella sampling algorithms to construct a potential of mean force[58]) or the transition end points (transition-path sampling[59]) is necessary.

Knowledge of thousands of high-resolution protein structures, together with the growing accessibility of various computational methods, has resulted in a large body of computational studies of protein dynamics. Given the power of computation, on the one hand, and the stringent prerequisite for accurate energetic descriptions of the system (small energy differences must be calculated relative to the absolute sum of all energetic terms of the system), on the other hand, experimental validation is necessary. Ideally, this should be an iterative process, with experimental testing of computational predictions and extensions of current computational methodology. This process is particularly important for tier-0 motions, because extensive approximations are required to gain access to this timescale computationally.

Fast timescales

'Fast' timescale dynamics (tier-1 and tier-2 dynamics) define fluctuations within the well of a tier-0 state. In contrast to the slow timescale, a large ensemble of structurally similar states that are separated by energy barriers of less than 1 kT result in more-local, small-amplitude picosecond-to-nanosecond fluctuations at physiological temperature (Fig. 1a). The interest in this timescale arises from the sampling of a large number of states, implicating these substates in the entropy of the system. In contrast to the tier-0 states, the large number of higher-tier states requires a statistical description of the distribution. We distinguish between tier-1 and tier-2 substates as small groups of atoms fluctuating collectively on the nanosecond timescale (such as loop motions) and local atomic fluctuations on the picosecond timescale (such as side-chain rotations), respectively (Fig. 1). We note that even higher tiers exist, such as femtosecond bond vibrations. Naturally, the structure dictates the features of atomic motion, with backbone atoms located in secondary structures being more restrained than atoms in loops. In this section, we discuss what has been learned about dynamics on the fast

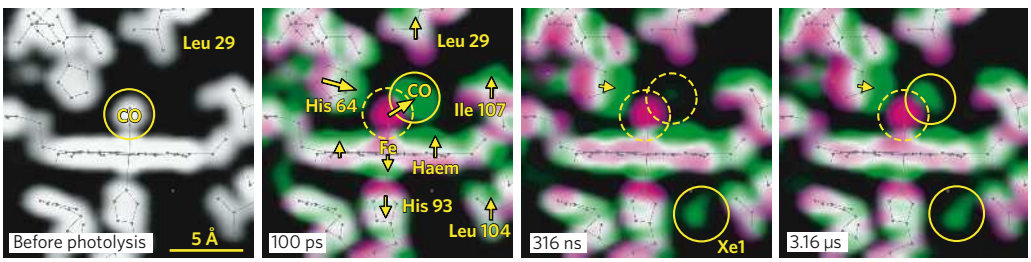


Figure 4 | Time dependence of carbon-monoxide migration and corresponding structural relaxation in myoglobin, using picosecond time-resolved X-ray crystallography. The ground-state electron density of carbon-monoxide (CO)-bound myoglobin is shown (left). Time-resolved changes after flash-photolysis-triggered dissociation of CO are displayed as coloured electron-density maps: the ground state is shown in pink, and the photolysed state in green; where these overlap, the colour blends to white. The direction of motion (indicated by arrows) follows the gradient from pink to green. Sites occupied by CO are indicated by solid circles, and sites evacuated by CO are indicated by dashed circles. The photolysed CO is initially trapped in the primary docking site, about 2 Å from the binding site and migrates subsequently to the xenon docking site Xe1 on the opposite side of the haem. Concurrent fast movements of active-site side chains (on the picosecond timescale) prevent immediate rebinding of CO. (Figure reproduced, with permission, from ref. 61.)

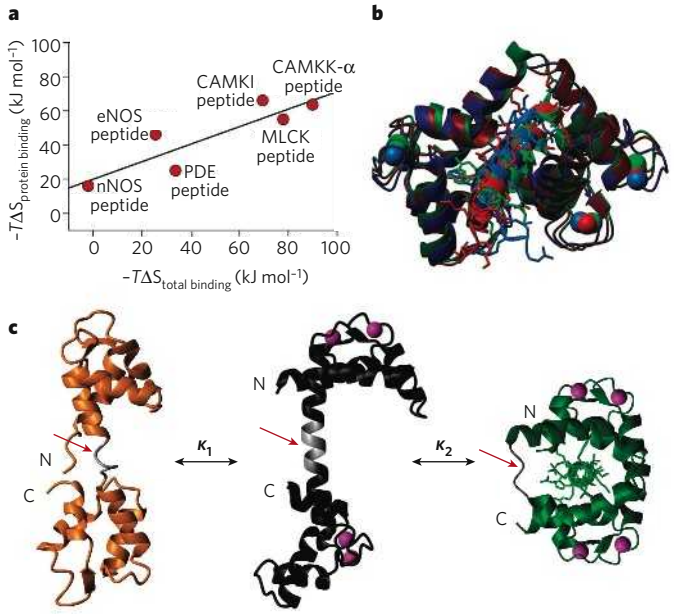


Figure 5 | The role of protein dynamics in molecular recognition by calmodulin on a range of timescales. a, Correlation between the change in conformational entropy of calmodulin ($T\Delta S_{protein\ binding}$) and the change in total system entropy ($T\Delta S_{total\ binding}$) on binding of peptides from target proteins. $T\Delta S_{protein\ binding}$ was estimated from methyl-NMR order parameters, and $T\Delta S_{total\ binding}$ was measured by isothermal titration calorimetry. (Panel reproduced, with permission, from ref. 72.) **b**, Overlay of X-ray crystal structures of calmodulin bound to several target peptides: calcium/calmodulin-dependent protein kinase I (CAMK1)-derived peptide (red); smooth-muscle myosin light-chain kinase (MLCK)-derived peptide (green); and endothelial nitric-oxide synthase (eNOS)-derived peptide (dark blue). Peptides bind in the centre of the structure (lighter shading). From this overlay, it is clear that there are large variations in the peptide side chains and consequently in the structure of calmodulin. Image generated from files from the PDB, based on data from the following: ref. 94, file 1MXE; ref. 95, file 1CDL; and ref. 96, file 1NIW. **c**, Mechanism of Ca^{2+} signalling and target recognition through coupled conformational equilibria. NMR experiments established that there is a dynamic equilibrium (K_1) between the structures of Ca^{2+}-free calmodulin (orange) and Ca^{2+}-bound calmodulin (black, Ca^{2+} in pink) in the absence of Ca^{2+}; Ca^{2+} binding then shifts this equilibrium to the right[74]. Further evidence from NMR spectroscopy indicates that the linker between the domains (grey, indicated by red arrows) remains flexible in the Ca^{2+}-bound state[75]. We propose that peptide binding occurs through selective binding to a pre-existing conformation of calmodulin (K_2 equilibrium), which is similar to the experimentally observed structure shown in green. Images generated from files from the PDB, based on data from the following: ref. 97, file 1CFD (left); ref. 76, file 1CLL (centre); and ref. 95, file 1CDL (right). CAMKK-α, calcium/calmodulin-dependent protein kinase kinase 1α; nNOS, neural nitric-oxide synthase; PDE, phosphodiesterase.

timescale from experimental atomic-resolution methods, experimental low-resolution and local-site methods, and computational methods.

Experimental atomic-resolution methods

X-ray-diffraction data contain information not only about the average three-dimensional structure (tier-0 state) but also about the spatial distribution around this state (tier-1 and tier-2 states). This mean-square atomic displacement[12] is commonly expressed as the B factor (also known as the temperature factor and the Debye–Waller factor). Atomic displacement can originate from both static disorder (that is, an ensemble of substates present in solution are trapped in the crystal) and dynamic disorder (that is, fluctuations that occur in the crystal). Thus, B factors cannot be interpreted simply as the amplitude of atomic fluctuations, because both true intramolecular motion and lattice disorder contribute to them. In addition, crystal contacts affect B factors. Recent advances in X-ray technology have resulted in structural models with sub-angstrom resolution, allowing novel insights into the directionality of atomic fluctuations through anisotropic B factors[60]. Laue X-ray diffraction can measure the purely dynamic component, with the added advantage of delivering the timescale of motions[19].

The elegance of Laue X-ray diffraction is illustrated by the time-resolved, high-resolution images that have been obtained for carbon-monoxide migration in myoglobin[61], which has been called the hydrogen atom of biology[62]. The dissociation of carbon monoxide from the haem cofactor was triggered by flash photolysis, and the resultant structural rearrangements were followed in real time[61] (Fig. 4). Clearly, correlated side-chain motions on the picosecond-to-nanosecond timescale coincide with carbon monoxide moving from its primary docking site into secondary pockets. This example highlights the role that angstrom-scale fast motions in the active site of myoglobin have in the reversible binding of the ligand, allowing a fast response to changes in the balance of oxygen and carbon monoxide in the blood. Unfortunately, this method cannot be universally applied to proteins, because the reaction needs to be triggered in the crystal and the structural changes must be small enough to be tolerated within the crystal lattice.

When using NMR relaxation methods, picosecond-to-nanosecond dynamics are characterized in terms of the amplitude (the order parameter, S^2, does not include directionality) and the timescale (τ_e, the internal correlation time) of bond fluctuations. S^2 ranges from 0 (isotropic rotation) to 1 (completely rigid) and is commonly measured for backbone amide bonds and side-chain methyl groups[24]. These local dynamics ('fast-timescale dynamics' in NMR-spectroscopy jargon) must be faster than the overall tumbling time of the protein to be detectable by solution NMR methods. In solid-state NMR spectroscopy, motions on a broader timescale (low microsecond and faster) can be detected[63,64]. In addition, there is no protein size limit in solid-state NMR spectroscopy, but technical challenges remain to be solved before dynamics can be routinely measured at atomic resolution in large proteins.

Because these fast-timescale dynamics are, ultimately, connected to entropy, NMR spectroscopy has been used extensively to investigate the entropic contribution of the protein to biomolecular binding (in protein–protein, protein–DNA, protein–RNA and protein–other-ligand interactions)[65–69]. Here, we use calmodulin to illustrate the advantages and limitations of this method for quantifying entropic contributions to affinity, because the natural function of calmodulin is to bind to a variety of target proteins. In the Ca^{2+}-bound form, the two domains clamp around the helical peptide-recognition sequence of target proteins (Fig. 5b, c). Although the free energy of binding is similar for Ca^{2+}-bound calmodulin interacting with many of these peptide sequences, the enthalpic and entropic contributions vary widely, as measured by isothermal titration calorimetry[70–72]. Using NMR spectroscopy, Wand and collaborators[72] identified differences in methyl order parameters of Ca^{2+}-bound calmodulin on binding to several peptide targets as contributing to this variation in entropy (Fig. 5a). However, the conversion of order parameters into absolute entropic energies is challenging and controversial[65–69,72,73]. A trend correlating the entropy calculated from the methyl-NMR order parameters of calmodulin and the total binding entropy of the system obtained from isothermal titration calorimetry is observed[72] (Fig. 5a), in agreement with the fact that many of these methyl groups line the peptide-binding pocket. Although the correlation is not strong, it is intriguing. This NMR method does not allow quantitative determination of the total protein entropy: only a subset of atoms is measured, and entropic changes in the peptides and solvent, which have not yet been characterized, might be major contributors to the total system entropy. The only component of the system that varies is the peptide itself, and a similar characterization of the peptides by NMR spectroscopy could be carried out. A comparison of known calmodulin–peptide structures exposes the differences in the calmodulin–peptide interfaces (Fig. 5b). Vastly different peptide side chains not only alter the packing of side chains with calmodulin but also displace the backbone of calmodulin. Packing shapes the amplitude and directionality of fluctuations, thus inspecting the structures should also allow insight into the entropic contribution to binding, as well as the enthalpic contribution of specific protein–peptide interactions.

The structural variation of these complexes raises the question of the mechanism of binding and specificity. The binding of Ca^{2+} to calmodulin was originally thought to induce activation through reorganization within each domain, but NMR relaxation experiments have revealed that the binding of Ca^{2+} shifts a pre-existing equilibrium[74]. We propose that peptide binding also proceeds through such an equilibrium-shift mechanism (Fig. 5c). This is supported by several lines of evidence: NMR data indicate that the linker helix between the domains is flexible in the absence of peptide[75]; X-ray diffraction has trapped both an extended structure (Protein Data Bank (PDB) identity 1CLL)[76] and a closed structure (PDB identity 1PRW)[77]; and single-molecule FRET distributions show that a wide range of interdomain distances are sampled[78,79]. This flexibility of calmodulin on the microsecond-to-millisecond timescale allows the ligands to select their preferred conformation, explaining the specificity of calmodulin for so many targets.

Experimental lower-resolution and local-site methods

Many of the low-resolution spectroscopic methods described in the section Slow timescales can also access this faster timescale. In addition, neutron scattering measures average root-mean-square fluctuations on any type of biological sample, thereby allowing dynamics to be characterized over a large temperature range. This method has demonstrated the influence of temperature and solvent conditions on the glass transition[80].

So far, we have discussed various timescales and their role in biological processes, but we have left out a fundamental motion: bond vibration on the femtosecond timescale. Advances in laser technology have initiated the fascinating era of femtosecond spectroscopy[81,82], and Zewail[82] and co-workers have developed the field of four-dimensional, ultrafast electron diffraction, crystallography and microscopy. These methods have extended the experimentally accessible time range, allowing direct observation of the basic chemical steps in enzymes: the breaking and forming of bonds, and the transfer of protons, hydride ions and electrons.

Computational methods

The most fundamental description of a system is computed using quantum mechanics, with molecular mechanics and molecular dynamics progressively simplifying the calculations to allow dynamics simulations on protein systems[2,47,49]. Tier-1 and tier-2 dynamics are on a perfect timescale for molecular-dynamics simulations. One of the advantages of molecular-dynamics simulations is that correlations between motions can be disentangled, a phenomenon that is obscured in experiments on ensembles. Most importantly, although experiments can determine what is moving and how fast, molecular-dynamics simulations can answer why things move, because the underlying forces and corresponding energies are included in the simulation. The resultant predictions inspire new experiments, forming part of a combined effort to solve the puzzle of how proteins work.

This point is exemplified by the problem of ion selectivity in potassium channels. How can a channel make a 'hole' in the membrane that allows ions to diffuse rapidly (10^8 ions per second) but discriminate 1,000-fold for K^+ over Na^+, which is only 0.4 Å smaller? The breakthrough crystal structure of a potassium channel, KcsA from *Streptomyces lividans* (PDB identity 1BL8)[83], provided unprecedentedly detailed structural information (Fig. 6a) and an immediate answer to this question. A narrow region of the pore, the selectivity filter, was perfectly sized to coordinate dehydrated K^+ but too large for dehydrated Na^+: the authors of this study concluded that "The structure of the selectivity filter with its molecular springs holding it open prevents the carbonyl oxygen atoms from approaching close enough to compensate for the cost of dehydration of a Na^+ ion"[83]. However, solution NMR spectroscopy shows increased flexibility in the selectivity filter relative to the transmembrane helices[84]. In addition, molecular-dynamics simulations on this structure of KcsA in a fully solvated lipid membrane show fluctuations of the selectivity filter over a range of ion–carbonyl distances that are sufficient to coordinate either ion[85] (Fig. 6b).

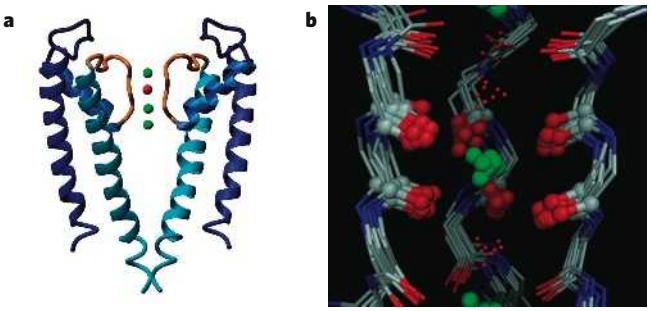


Figure 6 | Ion-channel selectivity investigated by X-ray crystallography and molecular-dynamics simulations. a, X-ray crystal structure of the potassium-selective channel KcsA. The structure has four subunits that create a central pore for ion conductance; for clarity, only two subunits are shown. The narrowest part of the pore was identified as the selectivity filter (orange). This region contains four ion-binding sites, in which backbone carbonyl groups coordinate K^+ ions perfectly. Three ions (green) and one water molecule (red) were observed in the selectivity filter. Image generated from file 1BL8 from the PDB, based on data from ref. 83. **b**, Superposition of snapshots of the selectivity filter from molecular-dynamics simulations. For clarity, only three subunits are shown. A K^+ ion (green) is coordinated in the S2 site by eight backbone carbonyl groups (six shown, with oxygen atoms depicted as large red spheres). Water molecules (small red circles) occupy adjacent sites. Atomic fluctuations of the selectivity filter on the order of 0.5–1 Å are captured in the molecular-dynamics trajectories. Nitrogen atoms are shown in blue, and carbon atoms are shown in grey. (Panel reproduced, with permission, from ref. 98.)

Importantly, this flexibility seems not to eliminate selectivity. The authors of the molecular-dynamics-simulation study[85] suggest that selectivity is controlled by the intrinsic electrostatic properties of the coordinating carbonyl groups and not by the average size of the pore measured by crystallography. Ultimately, selectivity is determined by the free-energy difference between K^+ and Na^+ partitioning between bulk water and the pore. The number of the coordinating groups, their nature and their distance distribution to the respective ion all contribute to this free-energy difference. To resolve the remaining controversies, the relative contributions of these factors need to be characterized precisely.

The hierarchy in space and time

In the previous sections, dynamics were separated into different timescales to discuss methodology, as well as examples that illustrate individual aspects of protein dynamics. However, a comprehensive description of the energy landscape requires connection between different timescales and the corresponding amplitudes of motions (Fig. 1a). Moreover, the ultimate goal is to understand how proteins function in real time. To live up to this task, several of the methods described here must be combined. This concept is illustrated using adenylate kinase (Fig. 7), an enzyme that catalyses the reversible conversion of an ATP and an AMP molecule into two ADP molecules.

Large conformational changes between the substrate-free enzyme and the substrate-bound enzyme have been observed[86] (Fig. 7a). Using ^{15}N-NMR relaxation experiments[34,39,40] on the turning-over enzyme, our research group showed that opening of the nucleotide 'lids', and not phosphotransfer, is the rate-limiting step for overall turnover[87]. Interestingly, comparison of the protein dynamics of a hyperthermophilic adenylate kinase (thermoAdk) and a mesophilic adenylate kinase (mesoAdk) showed that the reduced catalytic activity of thermoAdk at ambient temperature is solely due to slower lid opening[87].

X-ray structures of free and substrate-bound adenylate kinases suggest the standard view of ligand-induced conformational change[86]. However, the combined crystallographic, NMR spectroscopy, single-molecule and computational studies demand a fundamentally different picture[88]. The first hint came from observing three distinct conformations within the asymmetric unit of substrate-free thermoAdk (Fig. 7a). Remarkably, these trapped substates lie along the trajectory towards the closed

state. Because X-ray-crystallographic structures reveal high-resolution snapshots but do not provide the probability of sampling these states and other states, nor the rates of transition between them, we used NMR spectroscopy to measure the dynamics in solution. Indeed, collective conformational exchange with a common rate constant of about 1 ms was detected[88], but the data did not allow determination of the structural nature of the motion.

Molecular-dynamics simulations were therefore carried out for substrate-free Adk in explicit water (that is, individual water molecules were included in the simulation) to connect the spatial (X-ray) and kinetic (NMR spectroscopy) characteristics. The largest displacements occur in the nucleotide lids (Fig. 7b), and interconversion of the states observed in the three X-ray snapshots was reached within the 10 ns simulation time. If the fluctuations captured in the crystal happen on the nanosecond timescale, what process is detected on the millisecond timescale by NMR spectroscopy? Single-molecule FRET experiments on thermoAdk shed light on this question (Fig. 7d). Surprisingly, transitions between states that have dye–dye distances consistent with the fully open and fully closed states were detected for the free enzyme, highlighting the unique capability of single-molecule experiments to measure kinetics

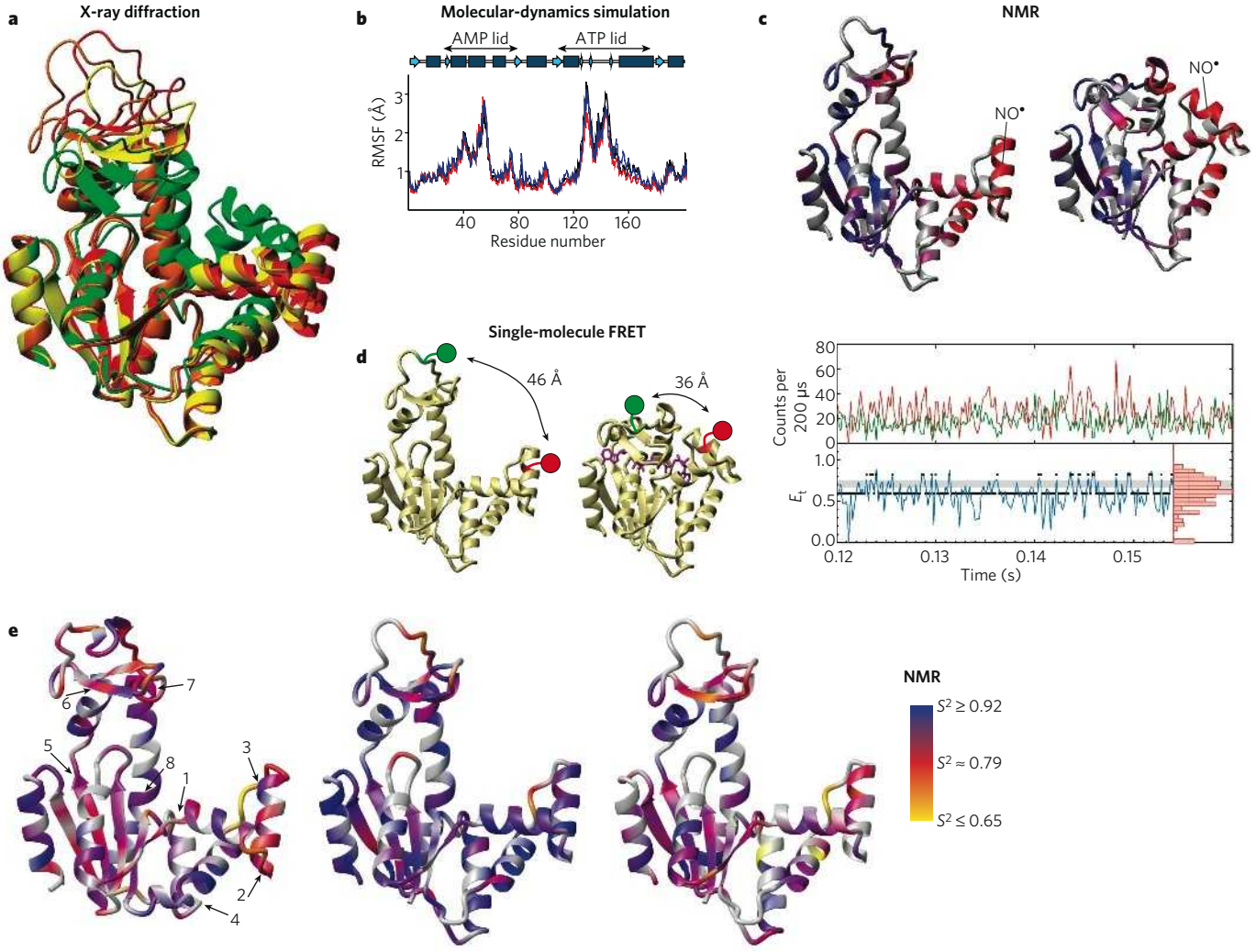


Figure 7 | A hierarchy of protein dynamics in space and time underlies enzyme catalysis, using the enzyme adenylate kinase as an example. a, The X-ray crystal structure of substrate-free thermoAdk captures snapshots along the trajectory towards the fully closed state. Molecules A, B and C in the asymmetric unit are shown in red, orange and yellow, respectively. The X-ray structure of thermoAdk bound to the bisubstrate analogue 5-di-adenosine-5′-pentaphosphate (green, substrate omitted for clarity) is superimposed. **b**, Root-mean-square fluctuations (RMSF) from 10-ns molecular-dynamics simulations of substrate-free thermoAdk molecules A, B and C (blue, red and black, respectively) show that the nucleotide lids are the most dynamic elements. A diagram of the secondary-structure elements is also shown, with α-helices indicated in dark blue and β-strands indicated in light blue (top). **c**, NMR paramagnetic relaxation enhancement (PRE) by a spin label (NO•) attached to residue 52. Substrate-free thermoAdk (left) samples conformations resembling the fully closed state (right). The PRE-derived distances for substrate-free thermoAdk are plotted onto the structures as a continuous colour scale from dark blue (distant, small effect) to red (close, large effect). Residues for which there are no data are shown in grey. **d**, Individual opening and closing events monitored by time-resolved single-molecule FRET of substrate-free immobilized thermoAdk. The positions of the FRET donor (green) and the FRET acceptor (red) on thermoAdk are indicated. The fluorescence intensities (upper panel of graph, green and red; colours correspond to donor and acceptor) are shown together with the corresponding FRET efficiencies (E_t, lower panel), including the E_t histogram (right). Each E_t value was assigned to either the open state (E_t below the grey band) or the closed state (E_t above the grey band) as indicated by black dots. **e**, NMR relaxation analysis of mesoAdk and thermoAdk. Fast (picosecond-to-nanosecond) atomic fluctuations are the physical origin of larger amplitude, slower nucleotide-lid motions. Order parameters (S^2) calculated from NMR relaxation data for mesoAdk at 20 °C (left) and thermoAdk at 20 °C (centre) and 80 °C (right) are shown as a continuous colour scale, with grey indicating proline residues and residues for which S^2 cannot be measured. The hinges are numbered and indicated with arrows. Importantly, at 20 °C, the picosecond-to-nanosecond hinge flexibility in mesoAdk is greater than in thermoAdk, and at this temperature, mesoAdk is known to be more active. In addition, the hinge flexibility on this timescale is similar in mesoAdk at 20 °C and thermoAdk at 80 °C, conditions in which both forms of the enzyme have similar activity. (Panel **a**–**d** reproduced, with permission, from ref. 88. Panel **e** reproduced, with permission, from ref. 89.)

when the spontaneous nature of the fluctuations impedes synchronization. Strikingly, the lifetime distributions of the open and closed states result in calculated rates that are in good agreement with the NMR rates measured for the ensemble. NMR paramagnetic relaxation enhancement, a powerful distance measure, unambiguously demonstrated sampling of a closed state. Severe line broadening was observed for residues that are far from the spin label in the open state but close to the spin label in the closed state (Fig. 7c). Thus, catalytically necessary conformational substates are already sampled in the free enzyme through motions with preferred directionality.

Turnover happens on the timescale of these tier-0, collective, large-amplitude motions. However, small-amplitude atomic thermal fluctuations occur on the picosecond timescale. How are these dynamic ranges connected? The link between these timescales was revealed by comparative analysis of thermoAdk and mesoAdk, using NMR spectroscopy and molecular-dynamics simulations[89]. Increased picosecond dynamics were observed in the same places where the local backbone conformation must change for lid closure to occur (Fig. 7e). Moreover, these hinges are more flexible in mesoAdk than thermoAdk at low temperature, with thermoAdk achieving similar hinge dynamics at temperatures at which the catalytic activity matches that of mesoAdk at ambient temperatures. This striking correspondence suggests that the physical origin of the catalytically important collective domain motions (microseconds to milliseconds) is the fast-timescale (picoseconds to nanoseconds) local hinge motions. Differences in the fast hinge fluctuations are encoded by differences in the amino-acid sequence, leading to increased packing and rigidification of thermoAdk on the picosecond-to-nanosecond timescale.

This example illustrates how the hierarchy of protein dynamics in space and time arises from the protein structure encoded by the amino-acid sequence and is ultimately connected to enzyme function. Tier-0 transitions are improbable, and therefore slow, events that arise from many individual attempts by local groups to overcome the energy barrier. The low success rate results from the collective nature of these large-scale motions.

From physics to biology and vice versa

From electrons and nuclei by way of X-rays, radio waves, light waves and energy potentials to a high-energy teenager surfing waves or the Internet — what is the connection? Biological function is ultimately rooted in the physical motions of biomolecules. Many biological processes are controlled by alterations in rates and relative populations rather than by a simple 'on–off' switch. For example, enzymes speed up chemical reactions, and changes in intracellular ion concentrations trigger complex neurological processes. Considering the immense rate enhancements and equilibrium shifts that are achieved in biological systems, it is easy to overlook the fact that only small changes in free energy (around a few kT) account for these effects, owing to the exponential dependence of both the rate and the populations on the free-energy difference. In other words, the breaking of a few hydrogen bonds or van der Waals contacts in a protein, which contains hundreds to thousands of such interactions, can turn on a signalling cascade or catalyse a chemical reaction. Importantly, intrinsic protein dynamics can happen only in this free-energy range of several kT.

Because biological function is the property selected by evolution, we propose that the conformational substates sampled by a protein, and the pathways between them, are not random but rather a result of the evolutionary selection of states that are needed for protein function. Signal transduction, enzyme catalysis and protein–ligand interactions occur as a result of the binding of specific ligands to complementary pre-existing states of a protein and the consequent shifts in the equilibria[26,35,36,75–79,90–92] (Fig. 1a). In other words, the dynamic landscape is an intrinsic property (or 'personality') of a protein and is encoded in its fold, and the ligand does not induce the formation of a new structure but, instead, selects a pre-existing structure.

The energy-landscape concept provides a vital bridge between the different philosophies and language used by physicists and biologists. For the field of biology to progress, quantitative analysis and the discovery of fundamental unifying principles, which are both characteristic of physics, are required. Conversely, the complexity of the living world provides a challenging task for physicists. It can only be imagined where this marriage of biology and physics might lead. Here, we pose a few immediate and well-defined questions about protein dynamics. How does a protein move from one energy valley into another — what is the pathway(s), and what is the transition state(s)? What are the entropic and enthalpic factors that contribute to transition barriers? Can minor conformational substates be predicted from known structures? Other important questions facing the field include how this knowledge can be used to design novel proteins that have desired properties and whether a dynamic view of proteins can be used to help discover and develop novel therapeutic agents.

Although there is certainly nothing wrong with having one eye on potential applications, many of the greatest advances in science have been unforeseen outcomes of basic discoveries, sparked solely by scientific curiosity. The beauty of scientific adventure is this unpredictable journey as a scientific community, following new instincts and evolving new directions. ■

1. Chandler, D. Roles of classical dynamics and quantum dynamics on activated processes occurring in liquids. *J. Stat. Phys.* **42,** 49–67 (1986).
2. Olsson, M. H. M., Parson, W. W. & Warshel, A. Dynamical contributions to enzyme catalysis: critical tests of a popular hypothesis. *Chem. Rev.* **106,** 1737–1756 (2006).
3. Benkovic, S. J. & Hammes-Schiffer, S. A perspective on enzyme catalysis. *Science* **301,** 1196–1202 (2003).
4. Gertner, B. J., Wilson, K. R. & Hynes, J. T. Nonequilibrium solvation effects on reaction-rates for model S_N2 reactions in water. *J. Chem. Phys.* **90,** 3537–3558 (1989).
5. Schliwa, M. (ed.) *Molecular Motors* (Wiley, Weinheim, 2003).
6. Kolomeisky, A. B. & Fisher, M. E. Molecular motors: a theorist's perspective. *Annu. Rev. Phys. Chem.* **58,** 675–695 (2007).
7. Lazaridis, T. & Karplus, M. 'New view' of protein folding reconciled with the old through multiple unfolding simulations. *Science* **278,** 1928–1931 (1997).
8. Leopold, P. E., Montal, M. & Onuchic, J. N. Protein folding funnels — a kinetic approach to the sequence structure relationship. *Proc. Natl Acad. Sci. USA* **89,** 8721–8725 (1992).
9. Wolynes, P. G. Recent successes of the energy landscape theory of protein folding and function. *Q. Rev. Biophys.* **38,** 405–410 (2005).
10. Austin, R. H., Beeson, K. W., Eisenstein, L., Frauenfelder, H. & Gunsalus, I. C. Dynamics of ligand binding to myoglobin. *Biochemistry* **14,** 5355–5373 (1975).
11. Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. The energy landscapes and motions of proteins. *Science* **254,** 1598–1603 (1991).
12. Frauenfelder, H., Petsko, G. A. & Tsernoglou, D. Temperature-dependent X-ray diffraction as a probe of protein structural dynamics. *Nature* **280,** 558–563 (1979).
13. Brooks, C. L. & Karplus, M. Solvent effects on protein motion and protein effects on solvent motion — dynamics of the active-site region of lysozyme. *J. Mol. Biol.* **208,** 159–181 (1989).
14. Fenimore, P. W., Frauenfelder, H., McMahon, B. H. & Parak, F. G. Slaving: solvent fluctuations dominate protein dynamics and functions. *Proc. Natl Acad. Sci. USA* **99,** 16047–16051 (2002).
15. Beece, D. *et al.* Solvent viscosity and protein dynamics. *Biochemistry* **19,** 5147–5157 (1980).
16. Fenimore, P. W., Frauenfelder, H., McMahon, B. H. & Young, R. D. Bulk-solvent and hydration-shell fluctuations, similar to α- and β-fluctuations in glasses, control protein motions and functions. *Proc. Natl Acad. Sci. USA* **101,** 14408–14413 (2004).
17. Shakhnovich, E. Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet. *Chem. Rev.* **106,** 1559–1588 (2006).
18. Lindorff-Larsen, K., Rogen, P., Paci, E., Vendruscolo, M. & Dobson, C. M. Protein folding and the organization of the protein topology universe. *Trends Biochem. Sci.* **30,** 13–19 (2005).
19. Bourgeois, D. & Royant, A. Advances in kinetic protein crystallography. *Curr. Opin. Struct. Biol.* **15,** 538–547 (2005).
20. Schlichting, I. *et al.* The catalytic pathway of cytochrome P450cam at atomic resolution. *Science* **287,** 1615–1622 (2000).
21. Englander, S. W. Hydrogen exchange and mass spectrometry: a historical perspective. *J. Am. Soc. Mass Spectrom.* **17,** 1481–1489 (2006).
22. Bai, Y. W. Protein folding pathways studied by pulsed- and native-state hydrogen exchange. *Chem. Rev.* **106,** 1757–1768 (2006).
23. Mittermaier, A. & Kay, L. E. New tools provide new insights in NMR studies of protein dynamics. *Science* **312,** 224–228 (2006).
24. Kay, L. E. NMR studies of protein structure and dynamics. *J. Magn. Reson.* **173,** 193–207 (2005).
25. Palmer, A. G. NMR characterization of the dynamics of biomacromolecules. *Chem. Rev.* **104,** 3623–3640 (2004).
26. Kern, D., Eisenmesser, E. Z. & Wolf-Watz, M. Enzyme dynamics during catalysis measured by NMR spectroscopy. *Methods Enzymol.* **394,** 507–524 (2005).
27. Pervushin, K., Riek, R., Wider, G. & Wuthrich, K. Attenuated T-2 relaxation by mutual cancellation of dipole–dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc. Natl Acad. Sci. USA* **94,** 12366–12371 (1997).
28. Sprangers, R., Gribun, A., Hwang, P. M., Houry, W. A. & Kay, L. E. Quantitative NMR spectroscopy of supramolecular complexes: dynamic side pores in ClpP are important for product release. *Proc. Natl Acad. Sci. USA* **102,** 16678–16683 (2005).
29. Palmer, A. G., Grey, M. J. & Wang, C. Y. Solution NMR spin relaxation methods for characterizing chemical exchange in high-molecular-weight systems. *Methods Enzymol.* **394,** 430–465 (2005).

30. Tugarinov, V. & Kay, L. E. Quantitative C-13 and H-2 NMR relaxation studies of the 723-residue enzyme malate synthase G reveal a dynamic binding interface. *Biochemistry* **44,** 15970–15977 (2005).
31. Sprangers, R. & Kay, L. E. Quantitative dynamics and binding studies of the 20S proteasome by NMR. *Nature* **445,** 618–622 (2007).
32. Horst, R. *et al.* Direct NMR observation of a substrate protein bound to the chaperonin GroEL. *Proc. Natl Acad. Sci. USA* **102,** 12748–12753 (2005).
33. Christodoulou, J. *et al.* Heteronuclear NMR investigations of dynamic regions of intact *Escherichia coli* ribosomes. *Proc. Natl Acad. Sci. USA* **101,** 10949–10954 (2004).
34. Loria, J. P., Rance, M. & Palmer, A. G. A relaxation-compensated Carr–Purcell–Meiboom–Gill sequence for characterizing chemical exchange by NMR spectroscopy. *J. Am. Chem. Soc.* **121,** 2331–2332 (1999).
35. Eisenmesser, E. Z., Bosco, D. A., Akke, M. & Kern, D. Enzyme dynamics during catalysis. *Science* **295,** 1520–1523 (2002).
36. Eisenmesser, E. Z. *et al.* Intrinsic dynamics of an enzyme underlies catalysis. *Nature* **438,** 117–121 (2005).
37. Fischer, G., Wittmannliebold, B., Lang, K., Kiefhaber, T. & Schmid, F. X. Cyclophilin and peptidyl-prolyl *cis–trans* isomerase are probably identical proteins. *Nature* **337,** 476–478 (1989).
38. Takahashi, N., Hayano, T. & Suzuki, M. Peptidyl-prolyl *cis–trans* isomerase is the cyclosporin-A-binding protein cyclophilin. *Nature* **337,** 473–475 (1989).
39. Loria, J. P., Rance, M. & Palmer, A. G. A TROSY CPMG sequence for characterizing chemical exchange in large proteins. *J. Biomol. NMR* **15,** 151–155 (1999).
40. Tollinger, M., Skrynnikov, N. R., Mulder, F. A. A., Forman-Kay, J. D. & Kay, L. E. Slow dynamics in folded and unfolded states of an SH3 domain. *J. Am. Chem. Soc.* **123,** 11341–11352 (2001).
41. Michalet, X., Weiss, S. & Jager, M. Single-molecule fluorescence studies of protein folding and conformational dynamics. *Chem. Rev.* **106,** 1785–1813 (2006).
42. Myong, S., Stevens, B. C. & Ha, T. Bridging conformational dynamics and function using single-molecule spectroscopy. *Structure* **14,** 633–643 (2006).
43. Yang, H. *et al.* Protein conformational dynamics probed by single-molecule electron transfer. *Science* **302,** 262–266 (2003).
44. Deniz, A. A., Mukhopadhyay, S. & Lemke, E. A. Single-molecule biophysics: at the interface of biology, physics and chemistry. *J. R. Soc. Interface* advance online publication, doi:10.1098/rsif.2007.1021 (22 May 2007).
45. Stryer, L. & Haugland, R. P. Energy transfer — a spectroscopic ruler. *Proc. Natl Acad. Sci. USA* **58,** 719–726 (1967).
46. Diez, M. *et al.* Proton-powered subunit rotation in single membrane-bound F_0F_1-ATP synthase. *Nature Struct. Mol. Biol.* **11,** 135–141 (2004).
47. Adcock, S. A. & McCammon, J. A. Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem. Rev.* **106,** 1589–1615 (2006).
48. McCammon, J. A., Gelin, B. R. & Karplus, M. Dynamics of folded proteins. *Nature* **267,** 585–590 (1977).
49. Scheraga, H. A., Khalili, M. & Liwo, A. Protein-folding dynamics: overview of molecular simulation techniques. *Annu. Rev. Phys. Chem.* **58,** 57–83 (2007).
50. Karplus, M. & Kushick, J. N. Method for estimating the configurational entropy of macromolecules. *Macromolecules* **14,** 325–332 (1981).
51. Ma, J. P. & Karplus, M. Ligand-induced conformational changes in *ras* p21: a normal mode and energy minimization analysis. *J. Mol. Biol.* **274,** 114–131 (1997).
52. Haliloglu, T., Bahar, I. & Erman, B. Gaussian dynamics of folded proteins. *Phys. Rev. Lett.* **79,** 3090–3093 (1997).
53. Jacobs, D. J., Rader, A. J., Kuhn, L. A. & Thorpe, M. F. Protein flexibility predictions using graph theory. *Proteins* **44,** 150–165 (2001).
54. Wells, S., Menor, S., Hespenheide, B. & Thorpe, M. F. Constrained geometric simulation of diffusive motion in proteins. *Phys. Biol.* **2,** S127–S136 (2005).
55. Hamelberg, D., Mongan, J. & McCammon, J. A. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *J. Chem. Phys.* **120,** 11919–11929 (2004).
56. Paci, E. & Karplus, M. Forced unfolding of fibronectin type 3 modules: an analysis by biased molecular dynamics simulations. *J. Mol. Biol.* **288,** 441–459 (1999).
57. Schlitter, J., Engels, M., Kruger, P., Jacoby, E. & Wollmer, A. Targeted molecular-dynamics simulation of conformational change — application to the T–R transition in insulin. *Mol. Simul.* **10,** 291–308 (1993).
58. Roux, B. The calculation of the potential of mean force using computer simulations. *Comput. Phys. Commun.* **91,** 275–282 (1995).
59. Dellago, C. & Bolhuis, P. G. Transition path sampling simulations of biological systems. *Top. Curr. Chem.* **268,** 291–317 (2007).
60. Merritt, E. A. Expanding the model: anisotropic displacement parameters in protein structure refinement. *Acta Crystallogr. D Biol. Crystallogr.* **55,** 1109–1117 (1999).
61. Schotte, F., Soman, J., Olson, J. S., Wulff, M. & Anfinrud, P. A. Picosecond time-resolved X-ray crystallography: probing protein function in real time. *J. Struct. Biol.* **147,** 235–246 (2004).
62. Frauenfelder, H., McMahon, B. H. & Fenimore, P. W. Myoglobin: the hydrogen atom of biology and a paradigm of complexity. *Proc. Natl Acad. Sci. USA* **100,** 8615–8617 (2003).
63. Franks, W. T. *et al.* Magic-angle spinning solid-state NMR spectroscopy of the β1 immunoglobulin binding domain of protein G (GB1): N-15 and C-13 chemical shift assignments and conformational analysis. *J. Am. Chem. Soc.* **127,** 12291–12305 (2005).
64. Lorieau, J. L. & McDermott, A. E. Conformational flexibility of a microcrystalline globular protein: order parameters by solid-state NMR spectroscopy. *J. Am. Chem. Soc.* **128,** 11505–11512 (2006).
65. Akke, M., Bruschweiler, R. & Palmer, A. G. NMR order parameters and free-energy — an analytical approach and its application to cooperative Ca^{2+} binding by calbindin-D_{9k}. *J. Am. Chem. Soc.* **115,** 9832–9833 (1993).
66. Jarymowycz, V. A. & Stone, M. J. Fast time scale dynamics of protein backbones: NMR relaxation methods, applications, and functional consequences. *Chem. Rev.* **106,** 1624–1671 (2006).
67. Lee, A. L., Sharp, K. A., Kranz, J. K., Song, X. J. & Wand, A. J. Temperature dependence of the internal dynamics of a calmodulin–peptide complex. *Biochemistry* **41,** 13814–13825 (2002).
68. Li, Z. G., Raychaudhuri, S. & Wand, A. J. Insights into the local residual entropy of proteins provided by NMR relaxation. *Protein Sci.* **5,** 2647–2650 (1996).
69. Yang, D. W. & Kay, L. E. Contributions to conformational entropy arising from bond vector fluctuations measured from NMR-derived order parameters: application to protein folding. *J. Mol. Biol.* **263,** 369–382 (1996).
70. Brokx, R. D., Lopez, M. M., Vogel, H. J. & Makhatadze, G. I. Energetics of target peptide binding by calmodulin reveals different modes of binding. *J. Biol. Chem.* **276,** 14083–14091 (2001).
71. Wintrode, P. L. & Privalov, P. L. Energetics of target peptide recognition by calmodulin: a calorimetric study. *J. Mol. Biol.* **266,** 1050–1062 (1997).
72. Frederick, K. K., Marlow, M. S., Valentine, K. G. & Wand, A. J. Conformational entropy in molecular recognition by proteins. *Nature* **448,** 325–329 (2007).
73. Best, R. B., Clarke, J. & Karplus, M. What contributions to protein side-chain dynamics are probed by NMR experiments? A molecular dynamics simulation analysis. *J. Mol. Biol.* **349,** 185–203 (2005).
74. Evenas, J., Forsen, S., Malmendal, A. & Akke, M. Backbone dynamics and energetics of a calmodulin domain mutant exchanging between closed and open conformations. *J. Mol. Biol.* **289,** 603–617 (1999).
75. Barbato, G., Ikura, M., Kay, L. E., Pastor, R. W. & Bax, A. Backbone dynamics of calmodulin studied by N-15 relaxation using inverse detected 2-dimensional NMR-spectroscopy — the central helix is flexible. *Biochemistry* **31,** 5269–5278 (1992).
76. Chattopadhyaya, R., Meador, W. E., Means, A. R. & Quiocho, F. A. Calmodulin structure refined at 1.7 angstrom resolution. *J. Mol. Biol.* **228,** 1177–1192 (1992).
77. Fallon, J. L. & Quiocho, F. A. A closed compact structure of native Ca^{2+}-calmodulin. *Structure* **11,** 1303–1307 (2003).
78. Johnson, C. K. Calmodulin, conformational states, and calcium signaling. A single-molecule perspective. *Biochemistry* **45,** 14233–14246 (2006).
79. Torok, K., Tzortzopoulos, A., Grabarek, Z., Best, S. L. & Thorogate, R. Dual effect of ATP in the activation mechanism of brain Ca^{2+}/calmodulin-dependent protein kinase II by Ca^{2+}/calmodulin. *Biochemistry* **40,** 14878–14890 (2001).
80. Doster, W., Cusack, S. & Petry, W. Dynamical transition of myoglobin revealed by inelastic neutron scattering. *Nature* **337,** 754–756 (1989).
81. Zhong, D. P. Ultrafast catalytic processes in enzymes. *Curr. Opin. Chem. Biol.* **11,** 174–181 (2007).
82. Zewail, A. H. 4D ultrafast electron diffraction, crystallography, and microscopy. *Annu. Rev. Phys. Chem.* **57,** 65–103 (2006).
83. Doyle, D. A. *et al.* The structure of the potassium channel: molecular basis of K^+ conduction and selectivity. *Science* **280,** 69–77 (1998).
84. Chill, J. H., Louis, J. M., Baber, J. L. & Bax, A. Measurement of ^{15}N relaxation in the detergent-solubilized tetrameric KcsA potassium channel. *J. Biomol. NMR* **36,** 123–136 (2006).
85. Noskov, S. Y., Berneche, S. & Roux, B. Control of ion selectivity in potassium channels by electrostatic and dynamic properties of carbonyl ligands. *Nature* **431,** 830–834 (2004).
86. Vonrhein, C., Schlauderer, G. J. & Schulz, G. E. Movie of the structural changes during a catalytic cycle of nucleoside monophosphate kinases. *Structure* **3,** 483–490 (1995).
87. Wolf-Watz, M. *et al.* Linkage between dynamics and catalysis in a thermophilic–mesophilic enzyme pair. *Nature Struct. Mol. Biol.* **11,** 945–949 (2004).
88. Henzler-Wildman, K. A. *et al.* Intrinsic motions along an enzymatic reaction trajectory. *Nature* **450,** 838–844 (2007).
89. Henzler-Wildman, K. A. *et al.* A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature* **450,** 913–916 (2007).
90. Volkman, B. F., Lipson, D., Wemmer, D. E. & Kern, D. Two-state allosteric behavior in a single-domain signaling protein. *Science* **291,** 2429–2433 (2001).
91. Tsai, C. J., Kumar, S., Ma, B. & Nussinov, R. Folding funnels, binding funnels, and protein function. *Protein Sci.* **8,** 1181–1190 (1999).
92. Boehr, D. D., Dyson, H. J. & Wright, P. E. An NMR perspective on enzyme dynamics. *Chem. Rev.* **106,** 3055–3079 (2006).
93. Ansari, A. *et al.* Protein states and protein quakes. *Proc. Natl Acad. Sci. USA* **82,** 5000–5004 (1985).
94. Clapperton, J. A., Martin, S. R., Smerdon, S. J., Gamblin, S. J. & Bayley, P. M. Structure of the complex of calmodulin with the target sequence of calmodulin-dependent protein kinase I: studies of the kinase activation mechanism. *Biochemistry* **41,** 14669–14679 (2002).
95. Meador, W. E., Means, A. R. & Quiocho, F. A. Target enzyme recognition by calmodulin — 2.4-angstrom structure of a calmodulin–peptide complex. *Science* **257,** 1251–1255 (1992).
96. Aoyagi, M., Arvai, A. S., Tainer, J. A. & Getzoff, E. D. Structural basis for endothelial nitric oxide synthase binding to calmodulin. *EMBO J.* **22,** 766–775 (2003).
97. Kuboniwa, H. *et al.* Solution structure of calcium-free calmodulin. *Nature Struct. Biol.* **2,** 768–776 (1995).
98. Noskov, S. Y. & Roux, B. Ion selectivity in potassium channels. *Biophys. Chem.* **124,** 279–291 (2006).

Acknowledgements We thank M. Börsch, P. Anfinrud and B. Roux for providing the original images of Figs 3, 4 and 6b.

Author information Reprints and permissions information is available at npg.nature.com/reprints. Correspondence should be addressed to D.K. (dkern@brandeis.edu).

The molecular sociology of the cell

Carol V. Robinson[1], Andrej Sali[2] & Wolfgang Baumeister[3]

Proteomic studies have yielded detailed lists of the proteins present in a cell. Comparatively little is known, however, about how these proteins interact and are spatially arranged within the 'functional modules' of the cell: that is, the 'molecular sociology' of the cell. This gap is now being bridged by using emerging experimental techniques, such as mass spectrometry of complexes and single-particle cryo-electron microscopy, to complement traditional biochemical and biophysical methods. With the development of integrative computational methods to exploit the data obtained, such hybrid approaches will uncover the molecular architectures, and perhaps even atomic models, of many protein complexes. With these structures in hand, researchers will be poised to use cryo-electron tomography to view protein complexes in action within cells, providing unprecedented insights into protein-interaction networks.

A cell consists of hundreds of different functional modules, such as the RNA exosome, the proteasome and the nuclear pore complex (NPC). These modules, in turn, are composed of macromolecules, such as proteins and nucleic acids, as well as various small molecules. 'Molecular sociology' refers to the interactions of molecules within these functional modules.

At one end of the scale, there are highly stable interactions that are robust enough to withstand the rigours of purification. A large proportion of these stable structures are likely to be solved. The preferred method for determining the structures of assemblies at atomic resolution is X-ray crystallography[1]. Crystallography, however, is suitable only for functional modules that can be reconstituted *in vitro* and purified in sufficient quantity for crystallization. A landmark in structural biology occurred in 2000, when atomic structures of a large functional module — the ribosome from extremophile bacteria — were solved[2–4]. Progress has since been made towards determining the structures of similarly large complexes; however, in the past decade, there has not been a marked increase in the molecular mass of asymmetrical complexes that can be studied by crystallography.

At the other end of the scale, there are interactions that occur more fleetingly, in response to intracellular signalling, for example. The potential for determining the structures of such transient complexes by using any type of crystallography is relatively poor. For these complexes, as well as for stable complexes that are refractory to structure determination by traditional methods, integrative approaches are required[5–8]. These approaches combine information from varied sources. For example, individual subunits can be assembled into the whole complex by molecular docking that is restrained by knowledge of structurally defined homologous interactions, direct contact information provided by mass spectrometry[9] and other data[10,11]. Such approaches have been aided greatly by the availability of high-resolution structures of individual subunits from high-throughput structural-genomics consortia[12], and they are enabling the generation of atomic models and architectural models (in which the location and orientation of subunits within an assembly are defined) of previously intractable assemblies[6,9,13,14]. These models provide a basis for the development of testable hypotheses that could not be envisaged without a structural model. A spectacular example of the use of a hybrid approach[15] is the molecular model of auxilin bound to clathrin (the main component of the coat of coated vesicles), which was obtained by fitting comparative protein-structure models of the components into a cryo-electron-microscopy map at 12 Å resolution[16] (Fig. 1). Difference mapping showed changes in the

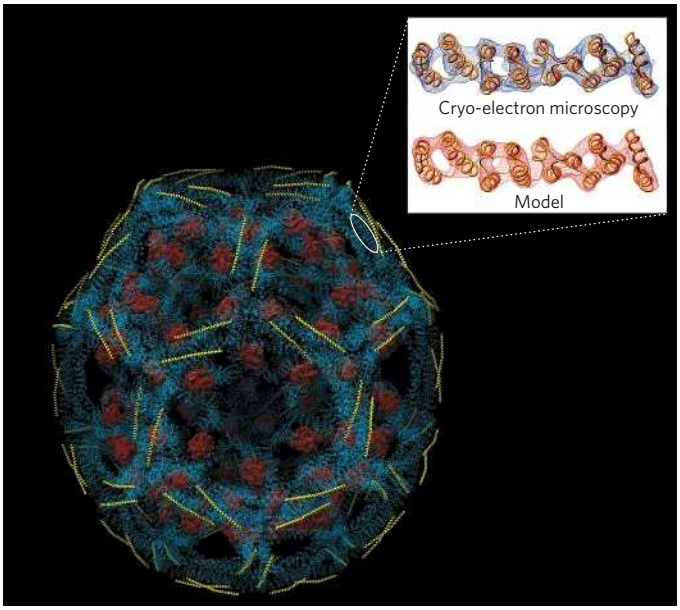


Figure 1 | A polypeptide-chain model for a clathrin D6 barrel. An α-carbon trace of the clathrin heavy (blue) and light (yellow) chains, derived by fitting atomic homology-based models into the density map from an 8 Å-resolution cryo-electron-microscopy reconstruction[16]. The position of a bound auxilin fragment (residues 547–910; red) was determined from a 12 Å-resolution cryo-electron-microscopy difference map. The inset zooms in to illustrate how closely the α-carbon coordinates of part of the heavy chain, as shown in the main figure (inset, lower), fit within the cryo-electron-microscopy density map (inset, upper). (Image reproduced, with permission, from ref. 16.)

[1]Department of Chemistry, Lensfield Road, University of Cambridge, Cambridge CB2 1EW, UK. [2]Department of Bioengineering and Therapeutic Sciences, Department of Pharmaceutical Chemistry, and California Institute for Quantitative Biosciences, Byers Hall, Suite 503B, University of California at San Francisco, 1700 4th Street, San Francisco, California 94158, USA. [3]Max Planck Institute of Biochemistry, Department of Molecular Structural Biology, Am Klopferspitz 18, D-82152 Martinsried, Germany.

clathrin lattice when auxilin is bound, prompting the hypothesis that local destabilization of the lattice promotes uncoating of the membranes of coated vesicles.

To illustrate the emergence of integrative approaches to structure determination, we have chosen a series of molecular 'machines' with differences in molecular mass, robustness and abundance: from the comparatively moderate dimensions of the yeast RNA exosome (400 kDa)[17] to the 26S proteasome (2.5 MDa)[18] and culminating in the NPC (50–100 MDa)[19]. For the yeast RNA exosome, which is relatively robust, atomic models were constructed by using spatial restraints from mass spectrometry[9] as a guide for the computational docking of subunit comparative models. By contrast, the heterogeneity and lability of the 26S proteasome have so far made it impossible to obtain a high-resolution model. The low resolution of the cryo-electron-microscopy map and the absence of high-resolution structures of many of the components — with the notable exception of the 20S core — have precluded the use of hybrid approaches to generate an atomic-resolution model. However, there are valuable data on binary interactions between the components, obtained from the yeast two-hybrid system and from mass spectrometry, and these need to be integrated with the cryo-electron-microscopy map. This example highlights the difficulties in applying integrative approaches to less-robust protein complexes. For the NPC, the highest-resolution *in situ* characterization was achieved recently by using cryo-electron tomography[20]. Moreover, it has also been possible to determine the configuration of the constituent proteins from a variety of proteomic and biophysical data[21]. Before presenting these examples, we consider the biophysical methods that can provide structural information about macromolecular assemblies.

Experimental methods for structure determination

Structures can be described at different levels of resolution. At the lowest level, the configuration of the components specifies the relative positions and interactions of the macromolecules. A higher-resolution description defines the molecular architecture, including the relative orientations of the components. For pseudo-atomic models, the positions of the atoms are specified but with errors larger than the size of an atom. The highest level of resolution is an atomic structure, which shows atomic positions with a precision smaller than the size of an atom.

Different experimental methods reveal different information about protein complexes. The stoichiometry and composition of an assembly, for example, can be determined by methods such as quantitative immunoblotting and mass spectrometry. The shape of the assembly can be revealed by cryo-electron microscopy and small-angle X-ray scattering (SAXS). In addition, cryo-electron microscopy can be used to determine the positions of the components, as can labelling techniques.

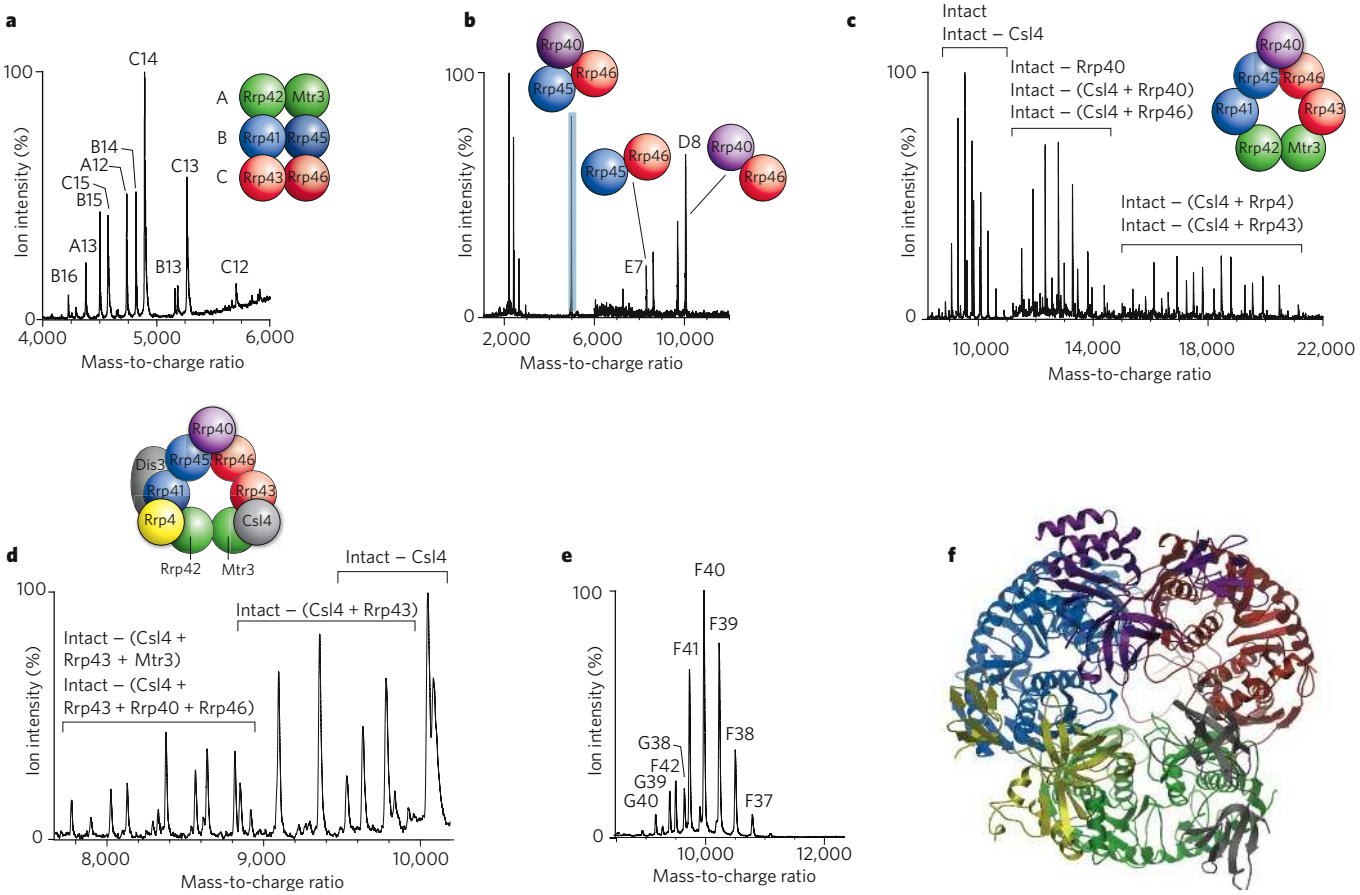


Figure 2 | Determining an atomic model of the yeast RNA exosome, by using mass spectrometry and comparative protein-structure modelling. The figure shows a series of five mass spectra, recorded under different conditions, revealing the building blocks from which the overall structure was constructed. **a**, Intact RNA exosomes were isolated from yeast and partially denatured. Mass spectrometry showed the presence of three heterodimers (A, B and C), as determined from the mass-to-charge ratio of each peak. (The number of positive charges corresponding to each dimer is indicated; for example, the largest peak represents the heterodimer C with 14 positive charges.) **b**, After tandem mass spectrometry (see page 991) of a low-abundance complex, highlighted in blue, that was present in the solution of the intact complex (not visible in **a**), a heterotrimer was identified that contained two of the subunits observed in (**a**) plus an additional subunit, Rrp40, enabling dimers B and C to be oriented within the ring. **c**, **d**, Using acceleration in the gas phase (**c**) and generation of complexes in solution (**d**), a series of related subcomplexes was produced, enabling the remaining subunits to be arranged in the ring, bridging subunits to be placed between the heterodimers, and the largest subunit, Dis3, to be located on the base of the complex. **e**, The intact complex confirms the single copy number of all ten subunits (F), with a small population of the complex having lost Csl4 during isolation (G). **f**, Comparative modelling was then used to produce an atomic model; the ribbons are depicted in colours corresponding to those in **a**–**d**. (Figure adapted, with permission, from ref. 9.)

Information can also be gained about whether particular components interact with each other, by using mass spectrometry, yeast two-hybrid experiments or affinity purification. Further information about interacting residues, as well as about the relative orientations of the components, can be inferred from cryo-electron microscopy, hydrogen–deuterium exchange, hydroxyl-radical footprinting and chemical crosslinking. At the highest resolution, information about the atomic structures of components and their interactions can be determined by X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. We outline some of these experimental methods in this section, and we highlight mass spectrometry and cryo-electron microscopy in the Boxes.

X-ray crystallography

The 'gold standard' for the structural analysis of proteins and protein complexes in terms of accuracy and resolution is X-ray crystallography[1]. Using this method, the amplitudes, and sometimes the phases, of structure factors in a crystal sample are measured. Together with a molecular-mechanics force field, this information is used in an optimization process that can result in an atomic structure of the assembly. In addition to the ribosome[2–4], X-ray crystallography has recently been used to solve structures of many macromolecular assemblies that involve protein–protein, protein–RNA and protein–DNA interactions, such as RNA polymerase[22], the RNA exosome[23] and the signal-recognition-particle complex[24].

NMR spectroscopy

NMR spectroscopy is increasingly used to determine which surfaces of components in a protein complex are interacting[25] (from chemical-shift perturbations[26] and residual dipolar coupling[27]), in addition to the structures of the individual protein components[28,29]. Such information can be combined with computational docking to obtain approximate structures of protein complexes[24]. A key attribute of NMR spectroscopy is that it allows the determination of atomic structures of complexes in solution in near-native conditions.

SAXS

Another method that enables structures to be determined in solution is SAXS. The data can be converted into a radial distribution function that provides low-resolution information about the shape of an assembly[30]. One of the advantages of SAXS is that it is suitable for assemblies of 50–250 kDa, which cannot easily be examined by cryo-electron microscopy or NMR spectroscopy. In addition, the ease of altering the solution conditions in which the sample is studied makes SAXS ideal for mapping differences between the conformational states of an assembly. The recent renaissance of SAXS largely results from efforts to integrate SAXS data with other structural information from complementary sources[31]. For example, the data obtained from SAXS studies of proteins or their complexes can be considered simultaneously with corresponding cryo-electron-microscopy maps[32]. SAXS spectra have also been incorporated into a protocol for structure determination by NMR spectroscopy[33]. Because SAXS data contain global information about the protein that is complementary to the short-range restraints from NMR spectroscopy, models of multidomain proteins are much more accurate than models based on NMR spectra alone. Examples of quaternary atomic structures obtained by using SAXS in conjunction with atomic structures of the protein components are calcium/calmodulin-dependent protein kinase II (ref. 34), the Ras activator son of sevenless (SOS)[35] and the various nucleotide-bound conformations of the ATPase GspE[36].

Labelling techniques

The approximate positions of protein components in an assembly can be determined by labelling techniques[37]. The protein component of interest is tagged with a probe, which can then be detected, for example by cryo-electron microscopy. The choice of labels depends on the known properties of the protein. For example, immuno-electron microscopy can be used to study proteins labelled with an antibody, which is typically conjugated to nanometre-sized gold beads to facilitate visualization[37]. Another option is to label protein components with histidine tags, which can be detected by using nickel-nitrilotriacetic acid (NiNTA)-conjugated gold beads[38]. Alternatively, proteins can be identified by exposing them to interacting proteins that have been covalently bound to gold beads[20].

Biochemical and biophysical methods

Information about the relative position, as well as the relative orientation, of the components in a complex can be gained from biochemical and biophysical methods. Site-directed mutagenesis, for example, can identify the amino-acid residues that mediate an interaction[14]. Approaching the same problem from a different angle, chemical footprinting[39] and hydrogen–deuterium exchange[40] can identify the surfaces that are buried when a complex forms[41]. Structural information

Box 1 | Mass spectrometry and hybrid approaches

Mass spectrometry has underpinned proteomics for many years, but recent developments have led to its integration into the structural-biologist's toolkit.

Determining the composition and stoichiometry of a complex

Analysis of intact complexes by mass spectrometry often requires modification of a conventional mass spectrometer, but it can reveal the stoichiometry and copy number of many protein complexes[58,82], from homomers (which consist of multiple copies of the same protein)[57] to complex heterogeneous structures such as ribosomes[83].

When using mass spectrometry as part of a hybrid approach, the first step (after the complex has been isolated by affinity purification) is a traditional proteomics experiment[84]. The component proteins of the complex are separated using one-dimensional SDS–polyacrylamide gel electrophoresis (SDS–PAGE) and are subjected to trypsin digestion. Mass spectra are then recorded for the resultant peptides. This first step, coupled with database searching to identify the peptides and therefore the subunits, provides an inventory of all components of the complex.

With the identity of the components established, the masses of the intact components can then be determined, to define post-translational modifications. This is achieved by using a denaturing step, typically incubation in a low-pH solution, to disassemble the complex — sometimes after chromatographic separation[45,85].

The next step is to generate subcomplexes by perturbing the complex in solution[9]. When the masses of the identified subunits and subcomplexes have been determined, a mass spectrum of the intact complex is used to define the overall copy number of all components in the complex. Computational analysis is then used to generate an interaction network[9]. The connections between each component are weighted according to the number of times each interaction is found in the population of networks that satisfy all of the nearest-neighbour restraints imposed by the subcomplexes.

Defining distance restraints by mass spectrometry

If two interacting proteins can be chemically crosslinked with molecular tethers, then this implies a distance restraint on the tethered residues[86]. Coupling this technique with mass spectrometry is appealing intellectually, but successes have so far been limited to a few small protein complexes[87]. Another way to define distance restraints is ion-mobility mass spectrometry[88]. This technique has only recently been applied to protein complexes, and, in these examples, measurement of collision cross-section has been used to examine complexes in the context of their X-ray structures[89]. One of the difficulties encountered when applying this approach is that protein complexes unfold when activated in the gas phase[90]. To restrain interaction models using collision cross-sections, it is necessary to establish conditions to maintain complexes as close to their native topology as possible. Although still in its infancy, ion-mobility mass spectrometry holds great promise for generating key structural information for the modelling of macromolecular assemblies.

can also be obtained by measuring the proximities of labelled groups on interacting proteins, using fluorescence resonance energy transfer (FRET) spectroscopy[42]. For example, the protein organization of the spindle pole body in yeast cells was established largely from distances obtained in FRET experiments[43].

Proteomics experiments

Proteomics experiments are generating large amounts of data that provide information about the molecular architectures of functional modules[6,7,43–45]. Information about binary interactions between proteins can be gained by using various techniques: yeast two-hybrid experiments[46,47], protein-fragment complementation assays[48], a combination of phage display and other techniques[49], protein arrays[50], and solid-phase detection by using surface plasmon resonance[51]. Physical interactions between proteins have also been inferred from genetic interactions, through the reduced activity or viability of mutant yeast strains in which genes encoding both proteins have been knocked out[52]. Furthermore, affinity purification[53,54] can be used to characterize not only binary interactions but also higher-order interactions, by purifying protein complexes and then identifying their components by mass spectrometry[55]; proximity between the identified components is established because they are directly or indirectly associated with the same tagged 'bait' protein.

Integration of structural information from different sources

After structural data have been obtained by one or more of these experimental methods, they need to be converted into a structural model through computation. As mentioned earlier, when approaches dominated by a single source of information fail, a hybrid approach, in which all of the available information about the composition and the structure of a given assembly is simultaneously considered (irrespective of the source), can sometimes be sufficient to calculate a useful structural model[5,6,8]. Even when this model is of relatively low resolution and accuracy, it can still be helpful for studying the function and evolution of the assembly; it also provides the necessary starting point for a study at higher resolution. An example of a simple hybrid approach is building a pseudo-atomic model of a large assembly by fitting the atomic structures of the subunits into the cryo-electron-microscopy map of the assembly[15,56,57]. In this section, we present three hybrid approaches, which were successfully applied to solve the structures of the RNA exosome, the 26S proteasome and the NPC.

One of the main difficulties encountered when structurally characterizing assemblies is the absence of information about direct contacts between subunits. Direct contacts can be identified by partial disruption of an assembly to yield a series of subcomplexes, followed by tandem mass spectrometry (which allows further disruption of a selected region of the mass spectrum) to determine the stoichiometry and the contacts between the components[58]. When enough subcomplexes have been characterized, an unequivocal protein–protein interaction network can be generated for the whole complex[7,9,45]. Such an approach has been applied to the yeast RNA exosome, which has ten subunits.

An atomic model of an RNA exosome

Despite its small size, attempts to analyse the eukaryotic RNA exosome by using X-ray crystallography have been repeatedly unsuccessful. Interesting structural insights have been gained, however, by overexpressing subunits of RNA exosomes from Archaea[59,60]. Moreover, a hybrid approach to studying the yeast RNA exosome has to some extent circumvented the challenges presented by crystallography. The yeast RNA exosome is present in both the nucleus and the cytoplasm, and is involved in RNA processing and turnover[17]. To obtain an architectural model of the yeast complex, the cytoplasmic form of the intact complex was isolated by tandem affinity purification[9]. Using partial denaturing agents, subcomplexes were generated and, after confirmation by tandem mass spectrometry, a protein–protein interaction network for the complex was determined (Fig. 2; Box 1). A key step in assembling the architectural model was the identification of three pairs of heterodimers that constitute a six-membered ring, a structure that had been observed in low-resolution electron-microscopy maps[61]. Experimental data also showed that several proteins — Csl4, Rrp4 and Rrp40 — bind to and

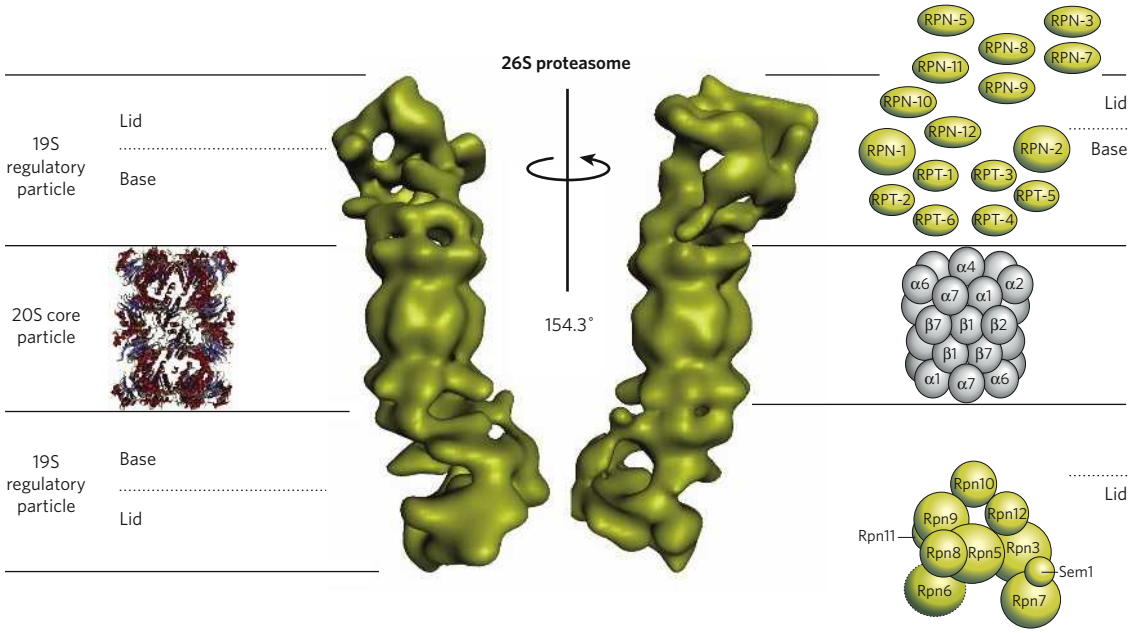


Figure 3 | The molecular architecture of the 26S proteasome. The 26S proteasome consists of 19S regulatory particles associated with the ends of a barrel-shaped 20S core particle. The part of each 19S regulatory subunit that is closest to the core is known as the base, and the part that is farthest away is known as the lid. Crystal structures have been obtained for archaeal, bacterial and eukaryotic 20S core particles[63,77–79] (left, α-helices in red, and β-sheets in blue). For the eukaryotic 26S holocomplex, only a low-resolution structure, obtained by cryo-electron microscopy[67], is available (centre; two orientations, rotated by 154.3°). Topological models of the regulatory particle have been deduced from yeast two-hybrid screens of *Caenorhabditis elegans* proteins[68] (upper right) and from mass spectrometry of yeast proteins[45] (lower right). These models agree reasonably well, albeit not completely. A topological model of the 20S core (centre right) that corresponds to the crystal structure (left) is also shown. No attempt has yet been made to obtain the molecular architecture of the entire 26S proteasome by integrating these topological models with the cryo-electron-microscopy map. RPN, non-ATPase subunit; RPT, ATPase subunit. (Central image reproduced, with permission, from ref. 65.)

strengthen the interfaces between the heterodimers, so these 'bridging' subunits were placed in the ring accordingly[9]. Given the similarity between the subunits in RNA exosomes from different species, models of the yeast proteins were then superimposed on the related archaeal ring structure[59]. The resultant model clearly shows the complementarity of the interactions within the various heterodimers and positions each of the bridging subunits between the heterodimers (Fig. 2). Restraints determined by mass spectrometry do not indicate whether the ring runs clockwise or anticlockwise, so the alternative enantiomer was also modelled. In this case, however, the interfaces within the heterodimers were less complementary than those in the first model, and the bridging subunits appear between the subunits within each heterodimer instead of between the heterodimers themselves. This arrangement is therefore not supported by experimental data on the bridging subunits[9]. Moreover, in this alternative model, the active sites of the catalytically active (RNase pleckstrin-homology (PH) domain) subunits — Rrp41 (also known as Ski6), Rrp46 and Mtr3 — are pointing towards the bridging subunits, which is in contrast to the known orientation of the Rrp41 equivalent in the archaeal RNA exosome[59]. An atomic model was then constructed (Fig. 2): this model is the best fit to the experimental data and is in close agreement with the structure of the related human RNA exosome, which was determined recently by using X-ray crystallography after reconstitution of nine subunits *in vitro*[23]. This example highlights the power of mass spectrometry and comparative protein-structure modelling to generate an atomic model of a complex protein assembly that has eluded determination by X-ray crystallography.

The architecture of the 26S proteasome

Determining the structure of the 26S proteasome presents an even greater challenge. Whereas the yeast RNA exosome can be isolated as a relatively homogeneous assembly, the 26S proteasome is labile and is therefore often heterogeneous. Moreover, unlike the yeast RNA exosome, there are few structures available for the components of the 26S proteasome, precluding atomic-resolution characterization.

The eukaryotic 26S proteasome is a large (2.5 MDa) molecular machine similar in size to the ribosome; it consists of one or two 19S regulatory complexes attached to the ends of a barrel-shaped 20S core complex. It has a central role in intracellular protein degradation, proteolytically cleaving proteins that have been marked for destruction by the attachment of multiple ubiquitin molecules[62]. The structure of the 20S core complex, which is highly conserved from Archaea to mammals, was solved by X-ray crystallography[63], revealing salient features of this protease[18]. A recent study also uncovered aspects of the structural changes that are involved in the functioning of the core complex, by using NMR spectroscopy[64]. By contrast, it has not been possible to crystallize the 26S holocomplex. The 19S regulatory subunits — which comprise at

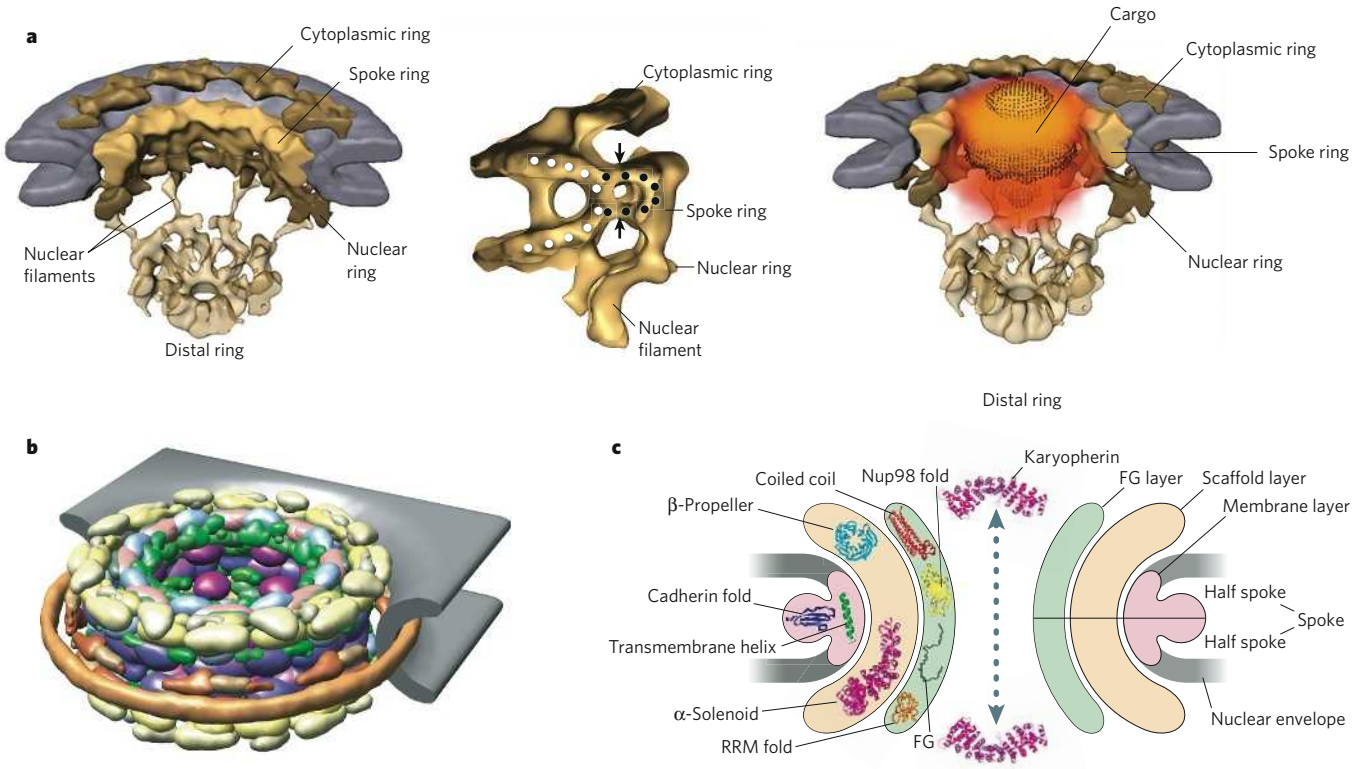


Figure 4 | The molecular architecture of the NPC. By using a variety of techniques, different aspects of the NPC structure have been revealed. **a**, Using cryo-electron tomography, a density map of the *Dictyostelium discoideum* NPC at 5.8 nm resolution was generated, allowing single molecules to be observed during nuclear import[20]. A cutaway view of the structure of rejoined asymmetrical units is shown (left), with subjective segmentation for the cytoplasmic ring, spoke ring and nuclear ring (brown and yellow), and the inner nuclear membrane and outer nuclear membrane (that is, the nuclear envelope; grey). For clarity, the central plug (that is, the transporter) has been omitted, and the basket with nuclear filaments and distal ring was rendered transparent. A cutaway view of a protomer is shown (centre). The fused inner nuclear membrane and outer nuclear membrane (white circles), as well as the clamp-shaped spoke structure (black circles), are indicated; arrows mark the entry and exit of what seems to be a channel. A cutaway view of the NPC structure with a three-dimensional probability distribution of import cargo is shown (right). The classical import cargo NLS–2GFP (Asn-Leu-Ser with two green fluorescent protein molecules attached) was labelled with gold, and the probability distribution for the cargo (orange; brightness indicates higher probability) is superimposed onto the central plug (brown dots). **b**, Various experimental data were integrated[7], revealing the configuration of the 456 core proteins (excluding FG (Phe-Gly) repeats in FG nucleoporins and the basket) that form the yeast NPC[21]. The inner and outer nuclear membranes (grey) are shown. The NPC proteins are coloured according to their assignment to various NPC modules: membrane rings (brown), outer rings (yellow), inner rings (purple, light and dark shades), linker nucleoporins (blue and pink, light shades) and FG nucleoporins (green). (Panel adapted, with permission, from ref. 7.) **c**, Structural folds were assigned to the domains of the NPC proteins, by comparing their sequences to those of known protein structures, revealing a simple fold composition and modular architecture for the NPC[72]. The architecture of the NPC ring, viewed as a transverse section, is segregated into three layers: membrane (pale pink), scaffold (pale yellow) and FG (pale green). The arrow denotes the direction of cargo transport. RRM, RNA-recognition motif.

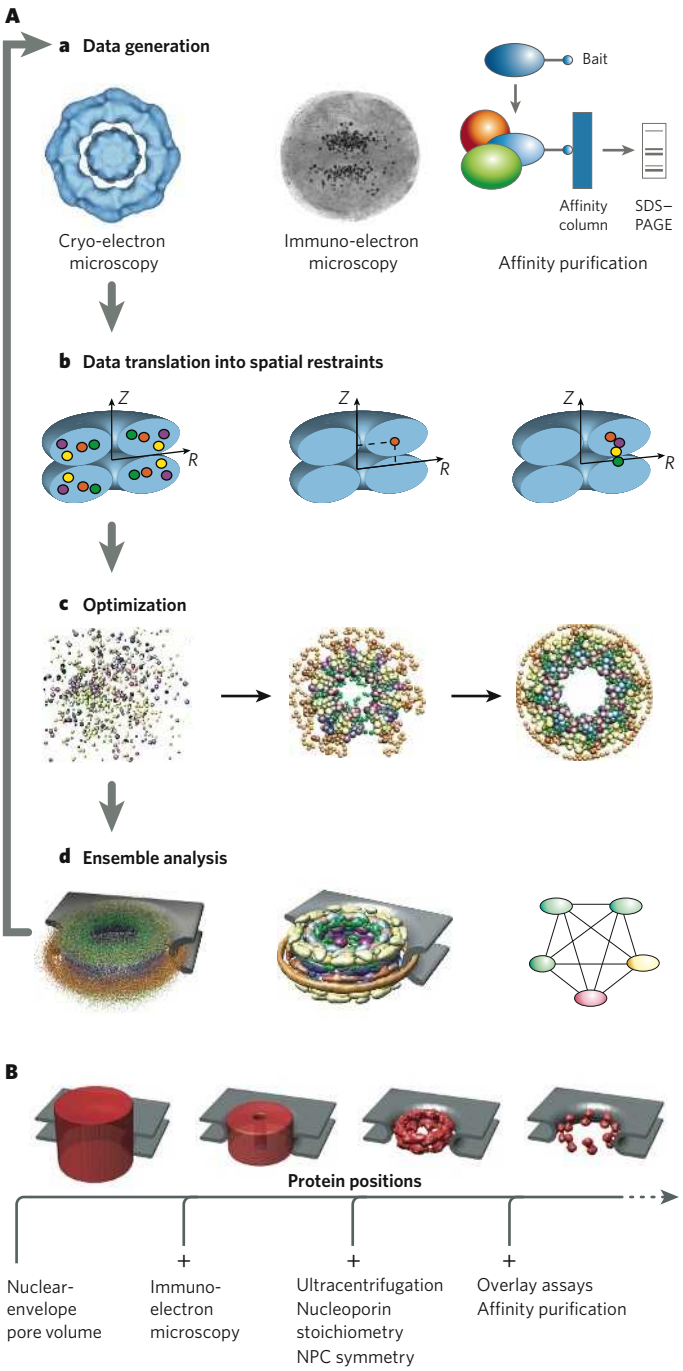


Figure 5 | Integrative structure determination. A, Using the NPC as an example[7], the four steps to determine a structure by integrating varied data are illustrated. These steps are data generation (**a**), data translation into spatial restraints (**b**), optimization (**c**) and ensemble analysis (**d**). **a**, First, structural data are generated by experiments, such as cryo-electron microscopy (left), immuno-electron microscopy (centre) and affinity purification of subcomplexes (right). Many other types of information can also be included. **b**, Second, the data and theoretical considerations are expressed as spatial restraints that ensure the observed symmetry and shape of the assembly (from cryo-electron microscopy, left), the positions of constituent gold-labelled proteins (from immuno-electron microscopy, centre) and the proximities of the constituent proteins (from affinity purification, right). The assembly is indicated in blue, and constituent proteins are indicated as coloured circles. **c**, Third, an ensemble of structural solutions that satisfy the data is obtained by minimizing the violations of the spatial restraints (from left to right). **d**, Fourth, the ensemble is clustered into sets of distinct solutions (left), and analysed in different representations, such as protein positions (centre) and protein–protein contacts (right). The integrative approach to structure determination has several advantages. First, synergy among the input data minimizes the drawbacks of incomplete, inaccurate and/or imprecise data sets. Each individual restraint contains little structural information, but by concurrently satisfying all restraints derived from independent experiments, the degeneracy of structural solutions can be markedly reduced. Second, this approach has the potential to produce all structures that are consistent with the data, not just one structure. Third, the variation between the structures that are consistent with the data allows an assessment of whether there are sufficient data and how precise the representative structure is. Last, this approach can make the process of structure determination more efficient, by indicating which measurements would be the most informative. **B**, When applying the process described in **A**, the position of each protein is specified with increasing accuracy and precision as each type of synergistic experimental information is added[7]. Each panel illustrates the localization volume (red) of 16 copies of nucleoporin 192 (Nup192) in the ensemble of NPC structures that satisfy the spatial restraints corresponding to the experimental data sets indicated. The smaller the volume, the better the proteins are localized. Further experiments could localize the proteins to a greater degree, as indicated by the dashed arrow. Therefore, the NPC structure is, in essence, 'moulded' into shape by the large quantity of diverse experimental data. (Panel reproduced with permission from ref. 7.)

least 18 subunits, including 6 ATPases — bind to ubiquitylated substrates and prepare them for degradation in the core complex. Structural studies of the 26S holocomplex, using cryo-electron microscopy, have been hampered by the low intrinsic stability of the complex, which tends to dissociate during purification and sample preparation. The dynamics of the complex present another problem: in addition to a set of 'canonical' subunits, there are several variable subunits; therefore, the composition of individual complexes varies, modulating proteasome function[65]. In principle, single-particle cryo-electron microscopy can handle heterogeneous samples that contain several distinct subsets of particles. Image classification allows particles to be sorted, thus achieving structural homogeneity *in silico*[66]. For a detailed classification, however, large sets of images are needed, and acquiring these is greatly facilitated by automated image recording[67]. At the present level of resolution (~2.5 nm), the spatial arrangement of the subunits of the 26S proteasome cannot be determined. Fortunately, there is a wealth of information on interactions between the proteasomal subunits, obtained from yeast two-hybrid studies[68] and mass spectrometry[45], as well as other sources[69] (Fig. 3). The challenge therefore is to interpret the current cryo-electron-microscopy map in light of these data. This should not be done in an *ad hoc* manner but by a systematic search for all structures that satisfy the restraints implied by the data. The power of such an approach is illustrated by the recent description of the architecture of the NPC[7,21].

The architecture of the NPC

NPCs are large proteinaceous assemblies that span the nuclear envelope, where they function as the main mediators of bidirectional exchange between the nucleoplasmic and cytoplasmic compartments in all eukaryotes[19]. Cryo-electron-microscopy images of the NPC show that it forms a channel through the stacking of two similar rings, each consisting of eight copies of the basic symmetry unit of the NPC (that is, the 'half spoke')[70]. In yeast, each half spoke contains ~30 different proteins known as nucleoporins, resulting in 456 proteins in the whole NPC, which has a mass of ~50 MDa[71]. Owing to its size and flexibility,

Box 2 | Cryo-electron microscopy

Cryo-electron microscopy is a generic term that refers to various electron-microscopy imaging modalities when applied to samples embedded in amorphous ice[91]. Samples are vitrified by plunge freezing or high-pressure freezing. A short description of the three main branches of cryo-electron microscopy is provided below.

Electron crystallography

Electron crystallography relies on the availability of two-dimensional crystals, either natural or synthetic. It is particularly suited to studying membrane proteins, but its use is not restricted to this class of protein. Very high resolution can be attained by optimizing the imaging conditions and by applying image-processing strategies to compensate for imperfections in the crystal lattices. Data acquisition can be time-consuming because of difficulties in collecting data sets of consistent quality; image quality is often degraded, particularly at high tilt angles, for reasons that are not well understood at present.

Single-particle analysis

Single-particle analysis (arguably a misleading name) relies on the existence of multiple copies of the object. Molecules suspended in thin layers of ice occur in random orientations. After grouping them into classes that correspond to common orientations, class averages are generated. Three-dimensional reconstructions are obtained by assigning relative orientations to the class averages and placing them in a virtual tilt experiment. Single-particle analysis is particularly suited to studying macromolecular complexes — the larger, the better. Some degree of heterogeneity in the sample (for example, variations in subunit composition, stoichiometry or conformational states) is tolerable and can be taken into account by image classification. There is no fundamental reason why atomic resolution could not be attained, but until now this has remained an elusive goal. Medium-resolution maps (1–2 nm) can be obtained routinely. This resolution is usually sufficient for fitting high-resolution structures of components (that is, subunits or domains) obtained by other methods into the cryo-electron-microscopy maps of the complex. At subnanometre resolution, elements of secondary structure can be discerned, enabling docking to be carried out with high accuracy. Efforts are under way to increase the speed of single-particle techniques, by automated data acquisition and image analysis[92].

Electron tomography

Electron tomography is unique in its capability to provide three-dimensional reconstructions of non-repetitive structures. Therefore, it enables insights into the molecular architecture of higher-order structures that have a degree of stochasticity. Objects are reconstructed from a series of transmission electron micrographs taken from different viewing angles. During data collection, the requirement for optimal sampling must be reconciled with the need to avoid radiation damage (through sustaining a low cumulative radiation dose). Tomograms taken in these conditions are rich in information, but the poor signal-to-noise ratio makes interpretation difficult. Tomograms of intact cells or organelles are images of their entire proteomes, and sophisticated pattern-recognition methods must be applied to make use of this information. At a resolution of 4–5 nm, typically obtained for intact cells, only large complexes can be visualized and mapped with an acceptable fidelity. With ongoing advances in instrumentation, however, resolutions of 2–3 nm are a realistic goal and will enable cells to be mapped more comprehensively. Better image-processing tools are needed to refine and validate such maps and to derive molecular-interaction patterns from them.

Generating pseudo-atomic models of assemblies

Fitting atomic structures and models of proteins and nucleic acids into cryo-electron-microscopy maps has resulted in pseudo-atomic models of many assemblies: complexes of viral subunits[93], ribosomes and ribosome-interacting proteins[94], the chaperone complex containing heat-shock protein 90 (ref. 95), cytoskeletal proteins and associated proteins[96,97], spliceosomal components[98], clathrin cages[16] and COPII cages[99]. Moreover, single-particle cryo-electron microscopy is becoming increasingly powerful at capturing assemblies in different conformational states[100].

detailed structural characterization of the complete NPC has proven to be extraordinarily difficult. Further compounding the problem, atomic structures have been solved only for domains that cover ~5% of the protein sequences[72]. As a result, the NPC is a challenging model system that is suitable for developing methods to map the molecular architectures of many other assemblies.

Cryo-electron tomography allows macromolecular assemblies to be studied *in situ*, eliminating the risk of preparation-induced artefacts and preserving the function of the structure[73] (discussed further in the next section). Thus, it is possible to take snapshots of molecular machines in action. This technique was applied to NPCs that were actively importing molecules into the intact nuclei of *Dictyostelium discoideum*. Many such snapshots were obtained and superimposed, yielding a map outlining the trajectories of the cargo[20] (Fig. 4a). Closer inspection of individually reconstructed NPCs shows substantial plasticity, probably reflecting both intrinsic dynamics and distortions that result from strain. To avoid the loss of resolution caused by averaging individually variable entities, a deformation analysis was carried out. This allows deviations from perfect eight-fold symmetry to be determined, and it provides the basis for the computational compensation of such distortions. Despite substantial improvements in resolution, the current resolution of 5.8 nm still falls short of that needed to determine the spatial arrangement of the component proteins.

The approximate spatial arrangement of the component proteins (Fig. 4b) can, however, be determined by integrating a variety of experimental data[7,21], using the approach outlined in Fig. 5. In a structure calculation, each of the 456 proteins in the yeast NPC was represented by a flexible chain consisting of a small number of connected beads (the numbers and radii of which were chosen to match the molecular masses and Stokes radii of the proteins). Next, to capture information about the structure of the NPC, a scoring function was constructed, which was a sum of spatial restraints of various types. These restraints incorporated data about protein shapes (from the protein sequences and ultracentrifugation), component protein positions (from immuno-electron microscopy), protein contacts (from affinity purification), eight-fold and two-fold symmetries of the NPC (from cryo-electron microscopy) and nuclear-envelope shape (from cryo-electron microscopy). The relative positions and proximities of the constituent proteins of the NPC were then calculated by satisfying these spatial restraints.

The calculation started with a random protein configuration and then iteratively moved the proteins so as to minimize violations of the restraints, relying on conjugate gradients and molecular dynamics with simulated annealing. To sample comprehensively all possible structural solutions that are consistent with the data, an 'ensemble' of 1,000 independently calculated structures that satisfy the input restraints was obtained. After superimposing the structures, the ensemble was converted into the probability of any volume element being occupied by a given protein (that is, the localization probability). The resultant localization probabilities yielded single pronounced maxima for almost all nucleoporins, showing that the input restraints define one predominant architecture for the NPC. The average standard deviation for the separation between nucleoporins is 5 nm. Given that this is less than the diameter of many NPC constituents, the map is sufficient to determine the relative positions of the proteins in the NPC. Although each individual restraint contains little structural information, the degeneracy of the structural solutions is markedly reduced by concurrently satisfying all restraints.

The arrangement of the proteins in the NPC (Fig. 4b, c), determined by the above approach, revealed that half of the NPC consists of a core scaffold, which is structurally analogous to vesicle-coating complexes[21,72]. This scaffold forms an interlaced network that coats the entire curved surface of the nuclear envelope, within which the NPC is embedded. The selective barrier to transport between the nucleoplasmic and cytoplasmic

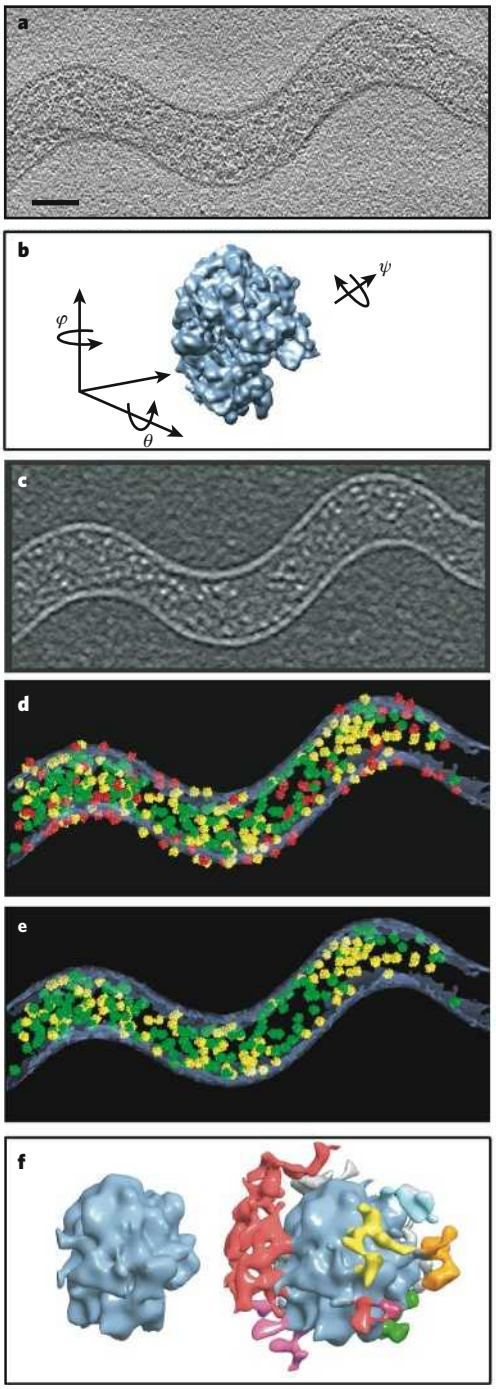


Figure 6 | Mapping of 70S ribosomes in a tomogram of the bacterium *Spiroplasma melliferum*[80]. a, An orthogonal slice through a tomogram of *S. melliferum* is shown. Scale bar, 100 nm. **b**, To determine the positions and orientations of the ribosomes in this cell, a template obtained by single-particle analysis[81] (resolution 11.5 Å) was correlated with the tomogram. **c**, In the cross-correlation function, white spots indicate sites where ribosomes were detected. **d**, From the cross-correlation function, a ribosome map was derived. Colours correspond to detection fidelity: high (green), intermediate (yellow) and low (red). **e**, After the initial ribosome map was generated, putative false positives were removed, leading to the refined map. The ribosomes that were identified and localized by template matching occupy ~5% of the cellular volume, which agrees well with estimates derived from other measurements. **f**, From the refined map, an average of the 70S ribosome was derived at a resolution of 45 Å (left). When the threshold for the isosurface representation of this map was lowered (right), distinct masses become visible near the ribosome. At present, these densities cannot be interpreted, but they most probably represent nascent chains, chaperones and other interacting factors. (Figure adapted, with permission, from ref. 80.)

compartments is formed by large numbers of FG nucleoporins, with disordered regions lining the inner face of the scaffold. The NPC consists of only a few structural modules. These modules resemble each other in terms of the configuration of their homologous constituents, thus providing clues to the ancient evolutionary origins of the NPC.

Studying functional modules *in situ*

Characterizing the NPC *in situ* required a non-invasive imaging technique. The technique used, cryo-electron tomography, generates images of large pleiomorphic objects — not only protein assemblies but also organelles. It does this by reconstructing three-dimensional objects from a series of two-dimensional transmission electron-microscopy images taken from different viewing angles.

Although the principles of electron tomography have been known for decades, its use has gathered momentum only recently. Technological advances have enabled the development of automated data-acquisition procedures, which in turn has reduced the total dose of electrons to a level at which radiation-sensitive biological materials, embedded in ice, can be studied[73] (Box 2). As a result, researchers are now poised to combine the potential of three-dimensional imaging with a 'close-to-life' preservation of biological specimens. At present, the resolution of cellular objects in cryo-electron-tomography studies is usually limited to 4–5 nm, but prospects for attaining molecular resolution (that is, 2–3 nm) are good[74].

Molecular-resolution tomograms of intact organelles or cells contain vast amounts of information. In essence, they are three-dimensional images of the entire proteome of a cell, and they should enable the spatial relationships of the macromolecules in a cell (the 'interactome') to be mapped (a process referred to as visual proteomics). Advanced pattern-recognition methods are needed to interpret the 'noisy' tomograms in an objective and systematic manner. This approach has two requirements: the proteomic 'inventory' must have been determined by mass-spectrometry analysis, and a library of template structures must be available so that tomograms can be interpreted by matching the cellular tomograms with the template structures[75]. Template structures can be generated by direct experimental methods, as well as by hybrid approaches. In the long term, with increasing numbers of structures of complexes deposited into the databases, template structures could be drawn from these databases.

We envisage a situation in which high-quality tomograms of a large range of cell types, generated with advanced instrumentation, will be made available to the scientific community, together with the software needed for their interpretation. This resource would enable researchers who have determined structures of complexes to use them as templates for exploring their functional environment. At the currently achievable resolution, only large complexes (such as ribosomes and proteasomes) can be mapped with an acceptable fidelity (Fig. 6; Box 2). But, with advances in instrumentation and methodology, today's imaging capabilities will improve, allowing proteomes to be mapped in a comprehensive manner. The remaining challenges are to untangle huge data sets, to derive interaction patterns from maps of intimidating complexity, and to understand the underlying molecular sociology.

Outlook

Constructing atomic models of functional modules in action will improve the current understanding of how cells function at many levels. To achieve this aim, new integrative methods are required, especially for dealing with the heterogeneity and dynamics of transient functional modules. One such hybrid approach that shows great promise is a combination of mass spectrometry and electron microscopy[76] in which isolation of functional modules is achieved in the gas phase. This allows selection of complexes on the basis of mass-to-charge ratio from a heterogeneous ensemble of closely related complexes. Subsequent 'soft landing' on suitable electron-microscopy grids then allows simultaneous characterization and visualization of transient complexes. These new hybrid methods, together with further computational integration, make revealing the molecular architecture of even fleeting social interactions within functional modules an enticing possibility. ■

1. Blundell, T. L. & Johnson, L. *Protein Crystallography* (Academic, New York, 1976).
2. Wimberley, B. T. *et al.* Structure of the 30S ribosomal subunit. *Nature* **407,** 327–339 (2000).
3. Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 Å. *Science* **289,** 905-920 (2000).
4. Schluenzen, F. *et al.* Structure of functionally activated small ribosomal subunit at 3.3 Å resolution. *Cell* **102,** 615–623 (2000).
5. Malhotra, A. & Harvey, S. C. A quantitative model of the *Escherichia coli* 16S RNA in the 30S ribosomal subunit. *J. Mol. Biol.* **240,** 308–340 (1994).
6. Alber, F., Kim, M. F. & Sali, A. Structural characterization of assemblies from overall shape and subcomplex compositions. *Structure* **13,** 435–445 (2005).
7. Alber, F. *et al.* Determining the architectures of macromolecular assemblies. *Nature* **450,** 683–694 (2007).
8. Sali, A., Glaeser, R., Earnest, T. & Baumeister, W. From words to literature in structural proteomics. *Nature* **422,** 216–225 (2003).
9. Hernandez, H., Dziembowski, A., Taverner, T., Seraphin, B. & Robinson, C. V. Subunit architecture of multimeric complexes isolated directly from cells. *EMBO Rep.* **7,** 605–610 (2006).
10. Davis, F. P. *et al.* Protein complex compositions predicted by structural similarity. *Nucleic Acids Res.* **34,** 2943–2952 (2006).
11. van Dijk, A. D. *et al.* Modeling protein–protein complexes involved in the cytochrome *c* oxidase copper-delivery pathway. *J. Proteome Res.* **6,** 1530–1539 (2007).
12. Todd, A. E., Marsden, R. L., Thornton, J. M. & Orengo, C. A. Progress of structural genomics initiatives: an analysis of solved target structures. *J. Mol. Biol.* **348,** 1235–1260 (2005).
13. Alber, F., Eswar, N. & Sali, A. in *Practical Bioinformatics* 1950–1954 (Springer, Heidelberg, 2004).
14. Sivasubramanian, A., Chao, G., Pressler, H. M., Wittrup, K. D. & Gray, J. J. Structural model of the mAb 806–EGFR complex using computational docking followed by computational and experimental mutagenesis. *Structure* **14,** 401–414 (2006).
15. Rossmann, M. G., Morais, M. C., Leiman, P. G. & Zhang, W. Combining X-ray crystallography and electron microscopy. *Structure* **13,** 355–362 (2005).
16. Fotin, A. *et al.* Structure of an auxilin-bound clathrin coat and its implications for the mechanism of uncoating. *Nature* **432,** 649–653 (2004).
17. Mitchell, P., Petfalski, E., Shevchenko, A., Mann, M. & Tollervey, D. The exosome: a conserved eukaryotic RNA processing complex containing multiple 3′→5′ exoribonucleases. *Cell* **91,** 457–466 (1997).
18. Baumeister, W., Walz, J., Zuhl, F. & Seemuller, E. The proteasome: paradigm of a self-compartmentalizing protease. *Cell* **92,** 367–380 (1998).
19. Lim, R. Y. & Fahrenkrog, B. The nuclear pore complex up close. *Curr. Opin. Cell Biol.* **18,** 342–347 (2006).
20. Beck, M., Lucic, V., Forster, F., Baumeister, W. & Medalia, O. Snapshots of nuclear pore complexes in action captured by cryo-electron tomography. *Nature* **449,** 611–615 (2007).
21. Alber, F. *et al.* The molecular architecture of the nuclear pore complex. *Nature* **450,** 695–701 (2007).
22. Meinhart, A. & Cramer, P. Recognition of RNA polymerase II carboxy-terminal domain by 3′-RNA-processing factors. *Nature* **430,** 223–226 (2004).
23. Liu, Q., Greimann, J. C. & Lima, C. D. Reconstitution, activities, and structure of the eukaryotic RNA exosome. *Cell* **127,** 1223–1237 (2006).
24. Egea, P. F. *et al.* Substrate twinning activates the signal recognition particle and its receptor. *Nature* **427,** 215–221 (2004).
25. Bonvin, A. M., Boelens, R. & Kaptein, R. NMR analysis of protein interactions. *Curr. Opin. Chem. Biol.* **9,** 501–508 (2005).
26. Zuiderweg, E. R. Mapping protein–protein interactions in solution by NMR spectroscopy. *Biochemistry* **41,** 1–7 (2002).
27. McCoy, M. A. & Wyss, D. F. Structures of protein–protein complexes are docked using only NMR restraints from residual dipolar coupling and chemical shift perturbations. *J. Am. Chem. Soc.* **124,** 2104–2105 (2002).
28. Wuthrich, K. The way to NMR structures of proteins. *Nature Struct. Biol.* **8,** 923–925 (2001).
29. Rieping, W., Habeck, M. & Nilges, M. Inferential structure determination. *Science* **309,** 303–306 (2005).
30. Vachette, P., Koch, M. H. & Svergun, D. I. Looking behind the beamstop: X-ray solution scattering studies of structure and conformational changes of biological macromolecules. *Methods Enzymol.* **374,** 584–615 (2003).
31. Nagar, B. & Kuriyan, J. SAXS and the working protein. *Structure* **13,** 169–170 (2005).
32. Tidow, H. *et al.* Quaternary structures of tumor suppressor p53 and a specific p53 DNA complex. *Proc. Natl Acad. Sci. USA* **104,** 12324–12329 (2007).
33. Grishaev, A., Wu, J., Trewhella, J. & Bax, A. Refinement of multidomain protein structures by combination of solution small-angle X-ray scattering and NMR data. *J. Am. Chem. Soc.* **127,** 16621–16628 (2005).
34. Rosenberg, O. S., Deindl, S., Sung, R. J., Nairn, A. C. & Kuriyan, J. Structure of the autoinhibited kinase domain of CaMKII and SAXS analysis of the holoenzyme. *Cell* **123,** 849–860 (2005).
35. Sondermann, H., Nagar, B., Bar-Sagi, D. & Kuriyan, J. Computational docking and solution X-ray scattering predict a membrane-interacting role for the histone domain of the Ras activator son of sevenless. *Proc. Natl Acad. Sci. USA* **102,** 16632–16637 (2005).
36. Yamagata, A. & Tainer, J. A. Hexameric structures of the archaeal secretion ATPase GspE and implications for a universal secretion mechanism. *EMBO J.* **26,** 878–890 (2007).
37. Hainfeld, J. F. & Powell, R. D. New frontiers in gold labeling. *J. Histochem. Cytochem.* **48,** 471–480 (2000).
38. Pye, V. E. *et al.* Structural insights into the p97–Ufd1–Npl4 complex. *Proc. Natl Acad. Sci. USA* **104,** 467–472 (2007).
39. Guan, J. Q., Almo, S. C., Reisler, E. & Chance, M. R. Structural reorganization of proteins revealed by radiolysis and mass spectrometry: G-actin solution structure is divalent cation dependent. *Biochemistry* **42,** 11992–12000 (2003).
40. Anand, G. S. *et al.* Identification of the protein kinase A regulatory $R^{I}\alpha$-catalytic subunit interface by amide $H/^{2}H$ exchange and protein docking. *Proc. Natl Acad. Sci. USA* **100,** 13264–13269 (2003).
41. Lee, T. *et al.* Docking motif interactions in MAP kinases revealed by hydrogen exchange mass spectrometry. *Mol. Cell* **14,** 43–55 (2004).
42. Yan, Y. & Marriott, G. Analysis of protein interactions using fluorescence technologies. *Curr. Opin. Chem. Biol.* **7,** 635–640 (2003).
43. Muller, E. G. *et al.* The organization of the core proteins of the yeast spindle pole body. *Mol. Biol. Cell* **16,** 3341–3352 (2005).
44. Gavin, A. C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440,** 631–636 (2006).
45. Sharon, M., Taverner, T., Ambroggio, X. I., Deshaies, R. J. & Robinson, C. V. Structural organization of the 19S proteasome lid: insights from MS of intact complexes. *PLoS Biol.* **4,** e267 (2006).
46. Parrish, J. R., Gulyas, K. D. & Finley, R. L. Yeast two-hybrid contributions to interactome mapping. *Curr. Opin. Biotechnol.* **17,** 387–393 (2006).
47. Uetz, P. *et al.* A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403,** 623–627 (2000).
48. Michnick, S. W., Ear, P. H., Manderson, E. N., Remy, I. & Stefan, E. Universal strategies in research and drug discovery based on protein-fragment complementation assays. *Nature Rev. Drug Discov.* **6,** 569–582 (2007).
49. Landgraf, C. *et al.* Protein interaction networks by proteome peptide scanning. *PLoS Biol.* **2,** e14 (2004).
50. MacBeath, G. & Schreiber, S. L. Printing proteins as microarrays for high-throughput function determination. *Science* **289,** 1760–1763 (2000).
51. Piehler, J. New methodologies for measuring protein interactions *in vivo* and *in vitro*. *Curr. Opin. Struct. Biol.* **15,** 4–14 (2005).
52. Collins, S. R. *et al.* Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* **446,** 806–810 (2007).
53. Krogan, N. J., Cagney, G., Haiyuan, Y., Zhong, G. & Guo, X. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440,** 637–643 (2006).
54. Collins, S. R. *et al.* Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics* **6,** 439–450 (2007).
55. Bauer, A. & Kuster, B. Affinity purification — mass spectrometry. Powerful tools for the characterization of protein complexes. *Eur. J. Biochem.* **270,** 570–578 (2003).
56. Rappas, M. *et al.* Structural insights into the activity of enhancer-binding proteins. *Science* **307,** 1972–1975 (2005).
57. Poliakov, A. *et al.* Macromolecular mass spectrometry and electron microscopy as complementary tools for investigation of the heterogeneity of bacteriophage portal assemblies. *J. Struct. Biol.* **157,** 371–383 (2007).
58. Hernandez, H. & Robinson, C. V. Determining the stoichiometry and interactions of macromolecular assemblies from mass spectrometry. *Nature Protoc.* **2,** 715–726 (2007).
59. Lorentzen, E. *et al.* The archaeal exosome core is a hexameric ring structure with three catalytic subunits. *Nature Struct. Mol. Biol.* **12,** 575–581 (2005).
60. Buttner, K., Wenig, K. & Hopfner, K. P. Structural framework for the mechanism of archaeal exosomes in RNA processing. *Mol. Cell* **20,** 461–471 (2005).
61. Aloy, P. *et al.* Structure-based assembly of protein complexes in yeast. *Science* **303,** 2026–2029 (2004).
62. Voges, D., Zwickl, P. & Baumeister, W. The 26S proteasome: a molecular machine designed for controlled proteolysis. *Annu. Rev. Biochem.* **68,** 1015–1068 (1999).
63. Groll, M. *et al.* Structure of 20S proteasome from yeast at 2.4 Å resolution. *Nature* **386,** 463–471 (1997).
64. Sprangers, R. & Kay, L. E. Quantitative dynamics and binding studies of the 20S proteasome by NMR. *Nature* **445,** 618–622 (2007).
65. Hanna, J. & Finley, D. A proteasome for all occasions. *FEBS Lett.* **581,** 2854–2861 (2007).
66. Scheres, S. H. W. *et al.* Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. *Nature Methods* **4,** 27–29 (2007).
67. Nickell, S. *et al.* Automated cryoelectron microscopy of 'single particles' applied to the 26S proteasome. *FEBS Lett.* **581,** 2751–2756 (2007).
68. Davy, A. *et al.* A protein–protein interaction map of the *Caenorhabditis elegans* 26S proteasome. *EMBO Rep.* **2,** 821–828 (2001).
69. Ferrell, K., Wilkinson, C. R., Dubiel, W. & Gordon, C. Regulatory subunit interactions of the 26S proteasome, a complex problem. *Trends Biochem. Sci.* **25,** 83–88 (2000).
70. Hinshaw, J. E., Carragher, B. O. & Milligan, R. A. Architecture and design of the nuclear pore complex. *Cell* **69,** 1133–1141 (1992).
71. Rout, M. P. *et al.* The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J. Cell Biol.* **148,** 635–651 (2000).
72. Devos, D. *et al.* Simple fold composition and modular architecture of the nuclear pore complex. *Proc. Natl Acad. Sci. USA* **103,** 2172–2177 (2006).
73. Koster, A. J. *et al.* Perspectives of molecular and cellular electron tomography. *J. Struct. Biol.* **120,** 276–308 (1997).
74. Nickell, S., Kofler, C., Leis, A. P. & Baumeister, W. A visual approach to proteomics. *Nature Rev. Mol. Cell. Biol.* **7,** 225–230 (2006).
75. Baumeister, W. From proteomic inventory to architecture. *FEBS Lett.* **579,** 933–937 (2005).
76. Benesch, J. L., Ruotolo, B. T., Simmons, D. A. & Robinson, C. V. Protein complexes in the gas phase: technology for structural genomics and proteomics. *Chem. Rev.* **107,** 3544–3567 (2007).
77. Lowe, J. *et al.* Crystal structure of the 20S proteasome from the archaeon *T. acidophilum* at 3.4 Å resolution. *Science* **268,** 533–539 (1995).
78. Unno, M. *et al.* The structure of the mammalian 20S proteasome at 2.75 Å resolution. *Structure* **10,** 609–618 (2002).
79. Kwon, Y. D., Nagy, I., Adams, P. D., Baumeister, W. & Jap, B. K. Crystal structures of the *Rhodococcus* proteasome with and without its pro-peptides: implications for the role of the pro-peptide in proteasome assembly. *J. Mol. Biol.* **335,** 233–245 (2004).
80. Ortiz, J. O., Forster, F., Kurner, J., Linaroudis, A. A. & Baumeister, W. Mapping 70S ribosomes in intact cells by cryoelectron tomography and pattern recognition. *J. Struct. Biol.* **156,** 334–341 (2006).
81. Gabashvili, I. S. *et al.* Solution structure of the *E. coli* 70S ribosome at 11.5 Å resolution. *Cell* **100,** 537–549 (2000).
82. Sharon, M. & Robinson, C. V. The role of mass spectrometry in structure elucidation of dynamic protein complexes. *Annu. Rev. Biochem.* **76,** 167–193 (2007).

83. Ilag, L. L. *et al.* Heptameric $(L12)_6$/L10 rather than canonical pentameric complexes are found by tandem MS of intact ribosomes from thermophilic bacteria. *Proc. Natl Acad Sci. USA* **102,** 8192–8197 (2005).
84. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422,** 198–207 (2003).
85. Synowsky, S. A., van den Heuvel, R. H., Mohammed, S., Pijnappel, P. W. & Heck, A. J. Probing genuine strong interactions and post-translational modifications in the heterogeneous yeast exosome protein complex. *Mol. Cell. Proteomics* **5,** 1581–1592 (2006).
86. Back, J. W., de Jong, L., Muijsers, A. O. & de Koster, C. G. Chemical cross-linking and mass spectrometry for protein structural modeling. *J. Mol. Biol.* **331,** 303–313 (2003).
87. Vasilescu, J. & Figeys, D. Mapping protein–protein interactions by mass spectrometry. *Curr. Opin. Biotechnol.* **17,** 394–399 (2006).
88. von Helden, G., Wyttenbach, T. & Bowers, M. T. Conformation of macromolecules in the gas phase: use of matrix-assisted laser desorption methods in ion chromatography. *Science* **267,** 1483–1485 (1995).
89. Ruotolo, B. T. *et al.* Evidence for macromolecular protein rings in the absence of bulk water. *Science* **310,** 1658–1661 (2005).
90. Ruotolo, B. T. *et al.* Ion mobility–mass spectrometry reveals long-lived, unfolded intermediates in the dissociation of protein complexes. *Angew. Chem. Int. Ed. Engl.* **46,** 8001–8004 (2007).
91. Henderson, R. Realizing the potential of electron cryo-microscopy. *Q. Rev. Biophys.* **37,** 3–13 (2004).
92. Suloway, C. *et al.* Automated molecular microscopy: the new Leginon system. *J. Struct. Biol.* **151,** 41–60 (2005).
93. Johnson, J. E. & Chiu, W. DNA packaging and delivery machines in tailed bacteriophages. *Curr. Opin. Struct. Biol.* **17,** 237–243 (2007).
94. Taylor, D. J. *et al.* Structures of modified eEF2 80S ribosome complexes reveal the role of GTP hydrolysis in translocation. *EMBO J.* **26,** 2421–2431 (2007).
95. Vaughan, C. K. *et al.* Structure of an Hsp90–Cdc37–Cdk4 complex. *Mol. Cell* **23,** 697–707 (2006).
96. Woodhead, J. L. *et al.* Atomic model of a myosin filament in the relaxed state. *Nature* **436,** 1195–1199 (2005).
97. Wang, H. W. & Nogales, E. Nucleotide-dependent bending flexibility of tubulin regulates microtubule assembly. *Nature* **435,** 911–915 (2005).
98. Stark, H. & Luhrmann, R. Cryo-electron microscopy of spliceosomal components. *Annu. Rev. Biophys. Biomol. Struct.* **35,** 435–457 (2006).
99. Fath, S., Mancias, J. D., Bi, X. & Goldberg, J. Structure and organization of coat proteins in the COPII cage. *Cell* **129,** 1325–1336 (2007).
100. Mitra, K. & Frank, J. Ribosome dynamics: insights from atomic structure modeling into cryo-electron microscopy maps. *Annu. Rev. Biophys. Biomol. Struct.* **35,** 299–317 (2006).

Acknowledgements We thank F. Alber, F. Foerster, M. Topf, D. Devos, J. Aitchison, C. Akey, M. Rout, B. Chait, R. Russell, H. Hernández, D. Matak-Vinkovic, M. Sharon, T. Taverner, J. Ortiz and S. Nickell. We also thank R. M. Glaeser for critical review of the manuscript. We are grateful to C. Johnson, S. Parker, C. Scheidegger and C. Silva of the Scientific Computing and Imaging Institute (University of Utah), and to R. K. Morley of RayScale, for help with preparing some of the images. We acknowledge funding from Interaction Proteome and 3D Repertoire (both funded by the European Commission), the Forum for European Structural Proteomics, the National Institutes of Health and the National Science Foundation.

Author information Reprints and permissions information is available at npg.nature.com/reprints. Correspondence should be addressed to the authors (cvr24@cam.ac.uk; sali@salilab.org; baumeist@biochem.mpg.de).

The origin of protein interactions and allostery in colocalization

John Kuriyan[1,2] & David Eisenberg[3]

Two fundamental principles can account for how regulated networks of interacting proteins originated in cells. These are the law of mass action, which holds that the binding of one molecule to another increases with concentration, and the fact that the colocalization of molecules vastly increases their local concentrations. It follows that colocalization can amplify the effect on one protein of random mutations in another protein and can therefore, through natural selection, lead to interactions between proteins and to a startling variety of complex allosteric controls. It also follows that allostery is common and that homologous proteins can have different allosteric mechanisms. Thus, the regulated protein networks of organisms seem to be the inevitable consequence of natural selection operating under physical laws.

One of the principal findings of molecular and cellular biology is that the metabolism and the homeostasis of cells are based on networks of interacting proteins[1]. Every protein interacts with other proteins: in some cases, by binding directly to another protein; in other cases, by modifying another protein; and in still other cases, by acting on a substrate and converting it into a substrate for the next protein in a pathway. Proteins that have crucial cellular roles are usually 'switchable', with their activities being modulated by other molecules. When such switching in protein activity is regulated by communication between two sites in a protein — the active site and the site of modification or binding — we refer to this as allostery. Allostery was defined originally as the regulation of a protein by a small molecule that differs in shape from the substrate, and this definition was later modified to the regulation of a protein through a change in its quaternary structure induced by a small molecule[2]. Our definition is broader than these and refers to a structural change — in the tertiary structure, the quaternary structure or both — induced by a small molecule or another protein. More generally, by our definition, the change induced by the modulator could be a change in the flexibility of the protein rather than simply a change in the structure[3]. In this broader sense, allostery accounts for the responsiveness of cells to external signals and for the regulation of metabolic pathways.

When confronting the intricacy of cellular networks and their exquisitely sensitive controls, scientists often wonder how such highly complex and regulated networks evolved. A few scientists go so far as to hold that "irreducibly complex" systems constitute a "powerful challenge to Darwinian evolution"[4]. The argument is that for a system that is "composed of several well-matched, interacting parts that contribute to the basic function, wherein the removal of any one of the parts causes the system to effectively cease functioning", these parts could not have evolved independently[4]. Protein networks with allosteric regulation are examples of such complex systems. Our view on the evolution of protein interactions and allostery is that natural processes of protein colocalization in cells, which effectively increase the local concentration of neighbouring molecules, change what might have seemed to be improbable evolutionary events into probable ones. This view complements ideas found in many earlier articles on this topic[2,5–8].

Fundamental 'forces', such as compartmentalization and electrostatic or hydrophobic binding, target proteins to specific locations in the cell, where they are colocalized with other proteins. This natural process of colocalization is essential in metabolism, transcriptional control, and signalling[9]. We argue that colocalization, combined with other natural processes (such as genetic recombination), leads naturally to protein complexes, to networks of interacting proteins and, subsequently, to allosteric control. Every protein complex or allosteric system that develops in this way might seem 'irreducibly complex', but these assemblies form as a result of the accidental mutations that first led to the interactions or fixed the relative disposition of the interacting domains. As a consequence, homologous proteins often have different allosteric mechanisms. Thus, although allostery is expected to arise naturally and readily in molecular 'machines', the precise mechanism is usually specific to one molecule and its closest relatives, and is not present across a protein family. This review explains how the regulated complexes and pathways of cells might have emerged, step by step, through natural selection working on proteins that have been colocalized by natural processes. First, we discuss how fundamental thermodynamic principles led to the idea that protein interactions and allostery emerge in a random manner as a consequence of colocalization. Then, we illustrate this principle with specific examples of diversity in the allosteric control mechanisms that govern homologous proteins.

The effect of colocalization

Some of the ways in which two proteins can be brought together in a cell are illustrated in Fig. 1. One way is the fusion of the genes that encode the two proteins so that the gene product now consists of two domains linked by a short segment of polypeptide chain. Such covalent linkage boosts the effective concentration of the protein domains with respect to each other to values in the range of 0.05–3.6 mM[10–13], greatly exceeding the usual

[1]Howard Hughes Medical Institute, California Institute for Quantitative Biosciences, Department of Molecular and Cell Biology and Department of Chemistry, University of California, Berkeley, California 94720, USA. [2]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. [3]Howard Hughes Medical Institute, University of California at Los Angeles–Department of Energy Institute of Genomics and Proteomics, Institute of Molecular Biology, Department of Biological Chemistry, and Department of Chemistry and Biochemistry, University of California, Los Angeles, California 90095, USA.

concentrations of proteins in cells, which tend to be in the nanomolar-to-micromolar range[14,15]. For example, a single molecule in a cell of the bacterium *Escherichia coli* has a concentration of ~1 nM.

The increased concentration that results from colocalization can have profound consequences for protein–protein interactions. In the highly atypical case of haemoglobin packed into erythrocytes, the concentration is ~5 mM[16]; this approaches the concentration (~12 mM) in the crystals that Max Perutz and co-workers used to determine the structure of haemoglobin[17]. In physiological conditions, however, the solubility of haemoglobin in erythrocytes is greater than its concentration, so haemoglobin does not precipitate or crystallize. By contrast, in individuals with the mutation that causes sickle-cell anaemia, in whom the glutamic-acid residue at position 6 of the β-chain of haemoglobin is replaced with a valine residue, the solubility of this mutant haemoglobin is half that of wild-type haemoglobin. The concentration of the mutant protein therefore exceeds its solubility, and it forms fibrils. This fibril formation distorts the erythrocytes, which are then poor hosts for the parasites that cause malaria. This process would not occur so readily without the high concentrations that result from the sequestration of haemoglobin in erythrocytes. Thus, in selecting the glutamic-acid-to-valine mutation, natural selection operates on the high concentration of colocalized haemoglobin molecules in erythrocytes.

Colocalization gives evolutionary processes the opportunity to convert nonspecific binding interactions into interactions that have functional consequences. Good examples of nonspecific binding interactions between protein molecules are the molecular contacts that occur within crystals in regions that are not part of a functional interface. Studies of these nonspecific contacts have revealed that they usually cover a small area, typically 200–1,200 Å^2 (ref. 18), and consist of a few hydrogen bonds and limited hydrophobic interactions. Such nonspecific interactions often occur when protein concentrations approach 1 mM.

The effect of colocalization on binding is large, as illustrated in Fig. 2. The binding curve shows, as a function of the concentration of protein A, the fraction (v) of another protein, B, that is bound to A. The fraction v is given approximately by the equation $v = ([A]/K_d)/(1 + [A]/K_d) = K_a[A]/(1 + K_a[A])$, where [A] is the concentration of A (strictly speaking, the activity of A), K_d is the dissociation constant for the binding of A to B, and K_a is the association constant for the binding of A to B (which equals $1/K_d$). Note that the binding of A to B is half complete ($v = 0.5$) when the concentration of A equals the dissociation constant. At a low concentration, such as $0.1K_d$, there is little binding, whereas at a high concentration, such as $10K_d$, nearly all of the binding sites on B are bound to A. Therefore, as the concentration of A varies from a low value typical of proteins in cells (~10–100 nM) to a value that can be achieved by colocalization within a cell (~1 mM), the nonspecific interaction between A and B can increase from negligible to substantial.

In the absence of colocalization, no single mutation is likely to convert a pair of proteins with a nonspecific binding affinity ($K_d > 10$ mM) to a binding pair. The reason is that a single residue change introduces no more than a few hydrogen bonds, each of which reduces the standard free energy of binding (ΔG^0) by ~1 kcal per mol[19]. A mutation that increases the nonpolar contact area could reduce the standard free energy of binding by no more than ~3 kcal per mol (for replacement of glycine with tryptophan)[20]. A reduction in the free energy of binding of ~3 kcal per mol reduces the dissociation constant by ~150-fold to ~0.1 mM (as calculated by $\Delta G^0 = RT \times \ln K_d = -RT \times \ln K_a$, where R is the

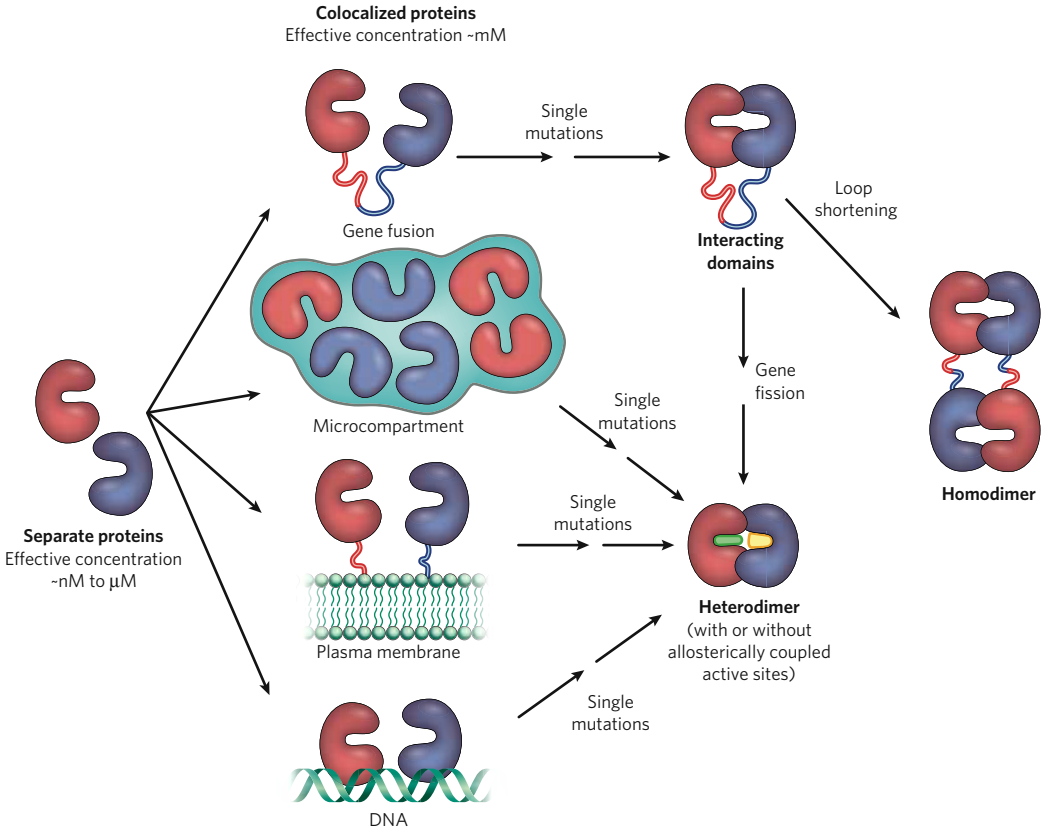


Figure 1 | The evolution of interacting proteins and allostery by single mutations. Two separate proteins in a cell are shown (left). Most cellular proteins are present at nanomolar-to-micromolar concentrations. A single random mutation in either protein is highly unlikely to result in binding or allostery. Interaction between these two proteins becomes probable when they are colocalized. Colocalization (second column) can occur by several mechanisms: by a gene fusion that results in both proteins being part of the same polypeptide chain, by concentration in a microcompartment, by association with the plasma membrane, or by binding to DNA. This process boosts the effective concentration of the proteins with respect to each other. Now, a single point mutation can lower the dissociation constant enough for a selectable change in function to occur. Further single mutations that increase the affinity of the two domains for each other, or that introduce allostery, can be selected for, resulting in tight interactions between these sites or allosteric coupling. Additional single genetic events such as gene fission or loop shortening can result in a strongly interacting heterodimer or an oligomeric homodimer.

gas constant and T is absolute temperature). Because the concentration of proteins in cells is usually in the nanomolar-to-micromolar range, binding is still negligible in this case, so no new complexes would form as a result of the mutation.

By contrast, when proteins are colocalized, a single mutation can lead to the formation of a new complex. The potential decrease in the dissociation constant caused by the new mutation could bring its value below the effective concentration (~1 mM) of the colocalized binding partner. The result would then be substantial binding, and the protein pair would constitute a complex with a new function. If the mutation increases the fitness of the organism, natural selection will fix it in the population. A series of such random mutations can lead, step by step, to a tighter binding site or to an interdependent set of allosteric interactions between the two domains (Fig. 1).

We therefore conclude that when two proteins are tethered to produce high effective concentrations, their colocalization greatly increases the probability that a random mutation in one of the proteins will change their mutual affinity and thus opens up the possibility of a change in fitness. Because the same process of colocalization followed by random mutation and natural selection can operate on assemblies of any number of component proteins, there is no reason to suppose that there is an 'edge' to the power of darwinian evolution beyond which the formation of complex biological structures must be attributed to 'deliberate intelligent design', as has been postulated[21]. The examples that follow suggest that complex networks of interacting proteins could indeed have evolved through processes of colocalization and that allosteric controls emerge by chance within these networks. Rather than presenting a paradox, the step-by-step evolution of complex, regulated networks emerges naturally when the laws of chemistry are coupled with natural selection.

Genetic fusion and the evolution of interacting proteins

A possible pathway for the evolution of a strongly interacting protein pair starting from two non-interacting proteins[22] is illustrated in the upper path of Fig. 1. When a random genetic event fuses the genes that encode two proteins, the expression product is a single polypeptide chain with domains corresponding to the initial proteins. Because these domains are colocalized on the same chain, their effective concentration increases from the nanomolar-to-micromolar concentration of the separated proteins to a millimolar concentration in the fused pair. Now, a single mutation in the gene can result in the replacement of an amino acid, possibly decreasing the free energy enough to create a tightly binding 'heterodimer' (still on the same chain) with an altered function, which is therefore selectable. Additional mutations can further stabilize the non-covalent bonds between the two domains or create allosteric interactions between them. At this point, another single genetic event can separate the gene that encodes the fused pair into two genes, each encoding one of the two proteins. Natural selection has therefore generated a heterodimer that can participate in signalling or metabolic processes.

Colocalization can also explain the evolution of multisubunit homo-oligomeric proteins[23]. The process can begin as shown in the upper pathway in Fig. 1. After the tightly binding multidomain protein has evolved, only a single genetic event is needed to convert the single-chain multidomain protein to a homodimer. This event is a genetic deletion that shortens the loop tethering the two original proteins. Such loop shortening can prevent the two domains from binding to one another but allow each domain to bind to its complementary domain in a second molecule. Numerous examples of oligomer formation by loop shortening have been observed in studies of genetically engineered proteins[24], strongly suggesting that such processes occur in nature.

Comparative genomics studies have found thousands of examples, over evolutionary timescales, of gene fusion resulting in a larger protein and of gene fission resulting in the constituent domains of a protein becoming separate proteins. These findings are collected in the protein-domain databases Pfam[25] and ProDom[26], and several examples are shown in Fig. 3. Predictions of pairs of proteins that bind to each other or participate in the same metabolic pathway can be made by finding a third protein that is homologous to both of the other proteins[22,27]. For example, it can be inferred that the α-subunit and the β-subunit of carbamoyl-phosphate synthetase from *E. coli* bind to each other because both are homologous to the longer sequence of the same enzyme in humans (Fig. 3). Indeed, in *E. coli*, these two subunits form a complex, the structure of which has been determined[28]. Systematic comparison of genomes shows that such gene-fusion and gene-fission events have been common in all three kingdoms of life[29–31].

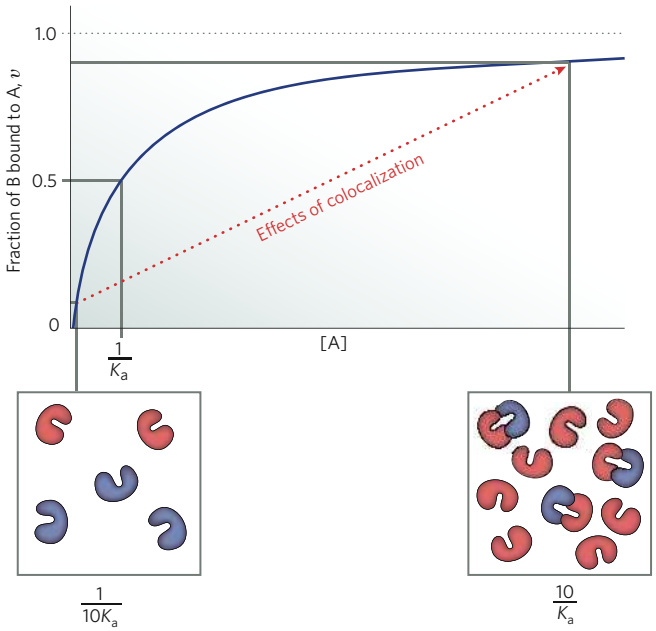


Figure 2 | The effect of colocalization on binding. Whether a protein, A (red), binds to another protein, B (blue), depends on the concentration of the first protein ([A]) and on the dissociation constant (K_d) of the complex (A–B). The fraction of B bound to A (v) is given approximately by the equation $v = ([A]/K_d)/(1 + [A]/K_d) = K_a[A]/(1 + K_a[A])$, where K_a is the association constant for the binding of A to B (which equals $1/K_d$). This describes a hyperbolic curve (blue). When $[A] = K_d = 1/K_a$, half of B is bound to A. A more exact relationship is needed, however, to distinguish between free A, plotted here on the x axis, and total A. This relationship gives a curve of a similar shape but with binding half-saturated at about $[A] = 0.4K_d$. Another approximation in this relationship is that activities are likely to differ from concentrations in the non-ideal environment of the crowded interior of a cell[67]. Despite these approximations, it remains true that when $[A] < 0.1K_d = 1/10K_a$, little A is bound to B (left inset), and when $[A] > 10K_d = 10/K_a$, B is nearly saturated with A (right inset). This means that colocalization, which boosts greatly the effective concentration of A (red arrow), results in increased binding. The grey dashed line indicates the asymptote.

In summary, commonly observed genetic events — recombinations, single-site mutations, and deletions — can account for the evolution of interacting proteins and of proteins in which multiple domains interact allosterically within a single chain. The allosteric mechanisms that emerge from colocalization and such genetic processes are likely to differ between homologous proteins, depending on which random mutations occurred in their evolutionary history. This principle is illustrated by the examples in the following three sections.

Allostery in DNA-binding proteins

Transcription factors recognize their target DNA sequences with high specificity by using cooperative binding: that is, the binding of one protein to DNA increases the affinity of another for an adjacent site[32,33]. These cooperative interactions between multiple domains extend the effective length of the target DNA sequence. This cooperative binding is important because a typical DNA-binding domain does not make contact with enough bases for the interaction to be highly specific. A striking example is the structure of the interferon-β enhanceosome,

in which eight proteins are bound to DNA adjacent to each other and, collectively, recognize an extended DNA element[34].

Cooperativity in DNA binding often results from the colocalization of DNA-binding domains in oligomeric assemblies or in a single polypeptide chain. For example, many DNA-binding proteins, such as λ repressor, are dimeric even when they are not bound to DNA[32]. Others, such as zinc-finger-containing proteins, have multiple DNA-binding domains in the same polypeptide chain[32]. In all such cases, the DNA-binding domains do not necessarily need to make contact with each other for the binding to be cooperative. There are many transcription factors, however, that interact with each other only when bound to DNA[33]. These interactions are allosteric, in the sense that contact with DNA results in the strengthening of protein–protein contacts. Although the mode of interaction with DNA is similar among evolutionarily related proteins that recognize DNA in this way, the nature of the protein–protein interactions can differ markedly. This principle is exemplified by the homeodomains, which are small DNA-binding domains found in many transcription factors that control development in animals.

The conserved core of homeodomains contains three α-helices, two of which form a helix–turn–helix motif[32]. By itself, a homeodomain binds weakly to DNA, but interactions with other DNA-binding domains, including other homeodomains, result in high-affinity and high-specificity DNA binding. In contrast to the highly conserved manner in which DNA is recognized by the homeodomain core, the interactions between the proteins bound to DNA are diverse, as shown for three homeodomain-containing proteins in Fig. 4a.

The homeodomain of *Drosophila melanogaster* Paired (PRD) proteins forms a highly cooperative homodimer on DNA, as a consequence of DNA deformations and reciprocal protein–protein contacts between the amino-terminal extension of one homeodomain and the second α-helix of the other[35] (Fig. 4a). The homeodomains of homeobox (HOX) proteins interact cooperatively with other homeodomains, such as those of *D. melanogaster* Extradenticle (EXD) and human pre-B-cell leukaemia homeobox 1 (PBX1). In contrast to the PRD dimer, in which the two homeodomains are located adjacent to each other on the DNA, the homeodomains of the HOX-protein–EXD and HOX-protein–PBX1 heterodimers are on opposite faces of the DNA, with the N-terminal linker of HOX reaching across the minor groove to engage the homeodomain of EXD or PBX1 (refs 36–38) (Fig. 4a).

Another distinct homeodomain interaction occurs for PIT1, which is a member of the POU family of transcription factors. These proteins contain a POU homeodomain and a POU-specific domain, the latter of which resembles the DNA-binding domains of bacterial transcription factors such as λ repressor[39]. PIT1 binds cooperatively to DNA as a dimer; the protein–protein interactions involve the DNA-recognition helix of the POU homeodomain of one PIT1 protein and a surface element of the POU-specific domain of the other[40] (Fig. 4a). This is in contrast to the situation for octamer-binding transcription factor 1 (OCT1), another member of the POU family, which recognizes DNA as a monomer[39]. Yet another mechanism is used by the homeodomain protein Matα2, which controls mating type in yeast (*Saccharomyces cerevisiae*). When Matα2 interacts with either of two proteins, Mat**a**1 or Mcm1, the affinity and the specificity of DNA binding increase. In the Matα2–Mat**a**1 complex, the extra carboxy-terminal helix present in the homeodomain of Matα2 engages the Mat**a**1 homeodomain[41]. By contrast, when the homeodomain of Matα2 interacts with Mcm1 (a MADS-box-domain-containing protein), the N-terminal region of the homeodomain mediates the contact[42].

The unrelated nature of these DNA-dependent protein–protein contacts suggests that they evolved after primordial homeodomains first developed affinity for DNA, with the spacing between the binding sites on DNA determining which regions of the proteins can make contact with each other. That is, colocalization seems to have preceded the different random mutations that produced different binding relationships in each of these cases.

Allostery in haemoglobin

The pressure differential between oxygen in the lungs and in the tissues (roughly threefold in humans) is too small for oxygen to be effectively transported to the tissues by simple diffusion. In humans, haemoglobin therefore evolved into a highly cooperative tetramer (consisting of two α-globin–β-globin heterodimers; Fig. 4b), which switches from a structure with low affinity for oxygen to one with high affinity, in an ultrasensitive manner[5]. A key feature of Perutz's classic mechanism of allostery in haemoglobin[5] is the coupling of the change in structure of an individual globin subunit after oxygen binding with the rotation of one α-globin–β-globin heterodimer with respect to the other. A crucial element in this coupling is the movement of the F helix, which is linked through a histidine residue to the iron in the haem group associated with each globin subunit[17].

All animals with blood-based respiratory systems need to cope with the limited pressure differential between points of oxygen uptake and release, so it is not surprising that allosteric haemoglobin molecules are common in animals. Genomic studies have revealed the striking conservation of globin subunits throughout evolution. But, in contrast to the conserved structure of globin subunits, biochemical and structural studies have shown that the mechanism discovered by Perutz does not hold for all allosteric haemoglobin molecules[43]. A haemoglobin of the clam *Scapharca inaequivalvis*, for example, binds to oxygen cooperatively but is a dimer rather than a tetramer[44] (Fig. 4b). The mechanism discovered by Perutz, which relies on rotation of one dimer in the tetramer with respect to the other, cannot operate in a dimeric haemoglobin. Instead, the allostery in the *S. inaequivalvis* haemoglobin results from a more direct transmission of the effects of oxygen binding at one haem group

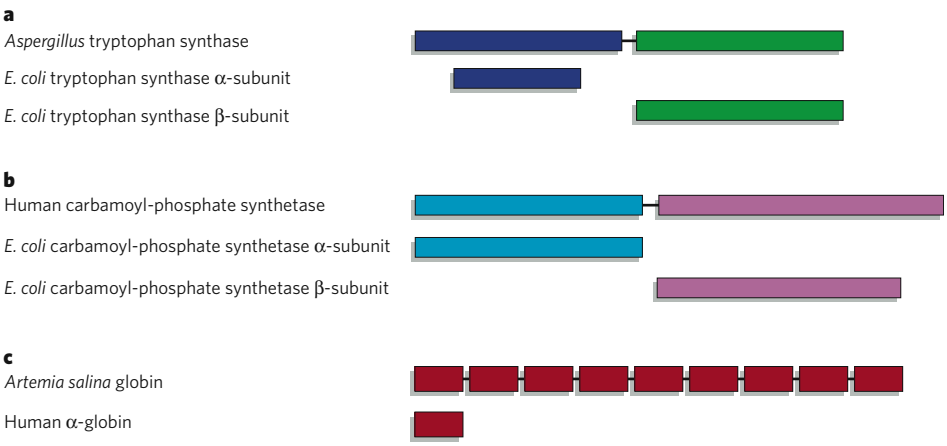


Figure 3 | Examples of fused protein domains in one organism that are homologous to separate domains in another organism. a, Tryptophan synthase. In *Escherichia coli*, the enzyme is a heterodimer formed from separate proteins that bind to one another. This pair of proteins is homologous to a single protein in *Aspergillus*. **b**, Carbamoyl-phosphate synthetase. The situation is similar to that shown in part **a**. **c**, Globin. One of the subunits of human haemoglobin, α-globin, is homologous to a globin protein from the brine shrimp *Artemia salina* that consists of nine fused globin domains.

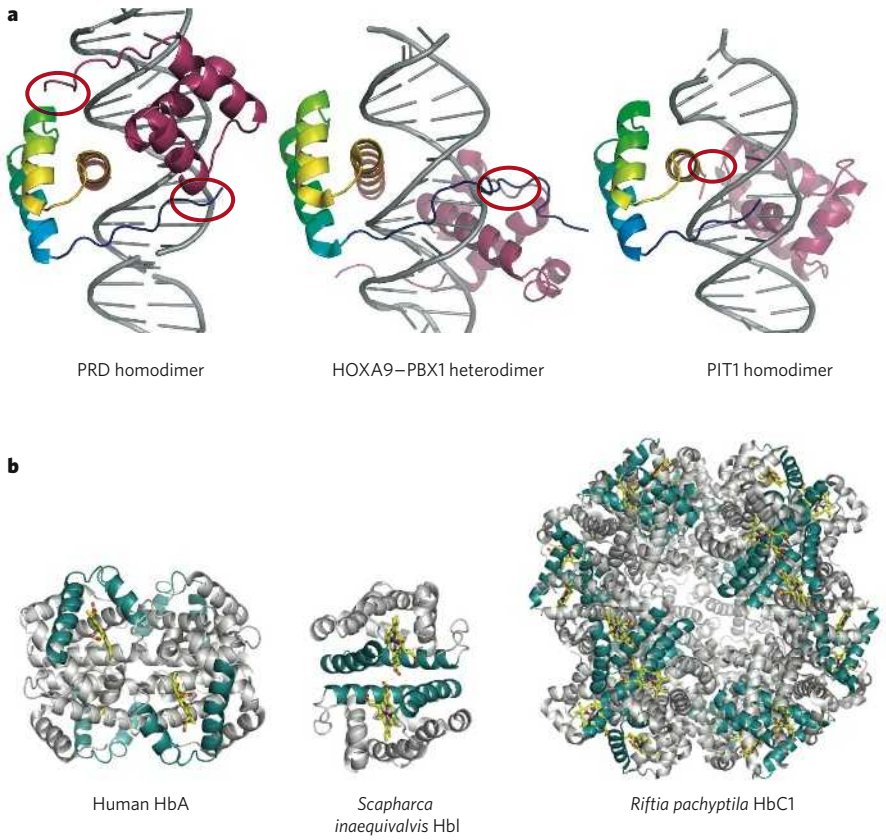


Figure 4 | Assemblies of homeodomain-containing proteins and haemoglobin. a, Three dimers of different proteins that contain homeodomains are shown bound to double-stranded DNA: a PRD homodimer, a HOXA9–PBX1 heterodimer, and a PIT1 homodimer. For clarity, the complete proteins are not shown. In each case, one homeodomain (left) is shown in the same orientation, with the colour varying from blue at the N terminus to pink at the C terminus. The domain that it interacts with is shown in pink, with the regions of contact indicated by red ovals. In the first two cases, the interaction occurs between homeodomains. For PIT1, the POU homeodomain in one molecule interacts with the POU-specific domain in the other molecule. Images generated from files from the Protein Data Bank (PDB), based on data from the following: ref. 35, file 1FJL (left); ref. 38, file 1PUF (centre); and ref. 40, file 1AU7 (right). **b**, Haemoglobin from three species is shown: humans (adult haemoglobin, HbA), the clam *Scapharca inaequivalvis* (HbI) and the tube worm *Riftia pachyptila* (HbC1). For each assembly, the two α-helices that bracket the haem group in each subunit are shown in blue. The haem group in each structure is shown in stick format, with carbon in yellow, oxygen in red, nitrogen in dark blue and iron in orange. Images generated from files from the PDB, based on data from the following: ref. 17, file 2HHB (left); ref. 68, file 4SDH (centre); and ref. 69, file 1YHU (right).

to the adjacent one, using conformational changes in the F helix that differ markedly from those in human haemoglobin.

A comparison of the quaternary structures of haemoglobin from various invertebrates reveals a striking diversity of assembly patterns[43]. Some haemoglobin molecules are dimeric; some are tetrameric; and others are organized into higher-order oligomers (Fig. 4b). The diversity of interfacial packing between globin subunits indicates that the allosteric mechanisms differ in each case. The linked network of molecular interactions seen in human haemoglobin — in which the change in size of the iron atom is transmitted through the iron-linked histidine residue and the F helix, causing breakage of ion pairs at the interfaces between subunits[5] — seems to be only one of several ways in which the globin fold can be adapted to yield a cooperative response to oxygen binding. Therefore, haemoglobin molecules in different organisms have acquired their allostery in several, apparently random, ways.

Mechanisms of control by phosphorylation

Phosphorylation is the most common covalent modification used to achieve allosteric control of proteins. In this section, we discuss two families of proteins — glycogen phosphorylases and a family of bacterial transcriptional activators — in which different mechanisms of control by phosphorylation have evolved within sets of homologous proteins. Then, we focus on protein tyrosine kinases, which carry out protein phosphorylation and are themselves regulated by phosphorylation, through extraordinarily diverse mechanisms.

Glycogen phosphorylases

Glycogen phosphorylase, which degrades glycogen to release glucose-1-phosphate, is activated by phosphorylation[45–47]. Glycogen phosphorylase is a dimeric enzyme, and the structures of the subunits and the general features of the dimeric assembly are conserved in the yeast and mammalian enzymes (Fig. 5a), as well as in related non-allosteric bacterial proteins[48]. The mechanism of control by phosphorylation of the yeast and mammalian enzymes is, however, different[48,49]. In both proteins, the site of regulatory phosphorylation is in segments at the N terminus of the polypeptide chain, but independent genetic-fusion events seem to have joined these unrelated segments to the conserved dimeric core of the enzyme[50,51]. In the yeast enzyme, the N-terminal tail of the protein blocks the active site, preventing access to substrates. Phosphorylation of a serine residue in the N-terminal tail causes removal of the tail from the active site, thereby activating the enzyme. By contrast, allosteric control of the mammalian enzymes is much more complex. Inactivation of the unphosphorylated protein results from distributed conformational changes, rather than from physical occlusion of the active site. The N-terminal residue that is phosphorylated in the yeast enzyme is not present in the mammalian enzyme. Instead, phosphorylation occurs at another site in the N-terminal region, and the phosphorylated segment is docked differently (Fig. 5a). The structural changes induced by phosphorylation include a rotation of one subunit with respect to the other, and these changes correlate with responsiveness to ATP (an inhibitor) and AMP (an activator). Neither of these molecules has an effect on the activity of the yeast enzyme.

Bacterial transcriptional activators of the AAA+ superfamily

Phosphorylation also regulates the function of a family of bacterial transcriptional activators that belongs to a superfamily of proteins (known as AAA+ ATPases) with diverse ATP-dependent functions. These transcriptional activators contain a signal-receiver domain that is controlled by phosphorylation and an ATPase domain that couples to the σ^{54}-containing form of RNA polymerase. The basic switch that controls the activity of these proteins is transition from a monomeric or dimeric form, both of which are inactive, to a ring-shaped assembly, typically a hexamer. This structure helps the σ^{54} subunit of RNA polymerase to unwind duplex DNA. Although the signal-receiver and ATPase domains are highly conserved within this family of transcriptional activators, the mechanism by which phosphorylation regulates activity is not[52–54] (Fig. 5b). In some members, such as nitrogen regulatory protein C (NtrC) from *Salmonella enterica* serovar Typhimurium, phosphorylation of the signal-receiver domain is required for the ATPase domain to oligomerize: that is, phosphorylation controls activity

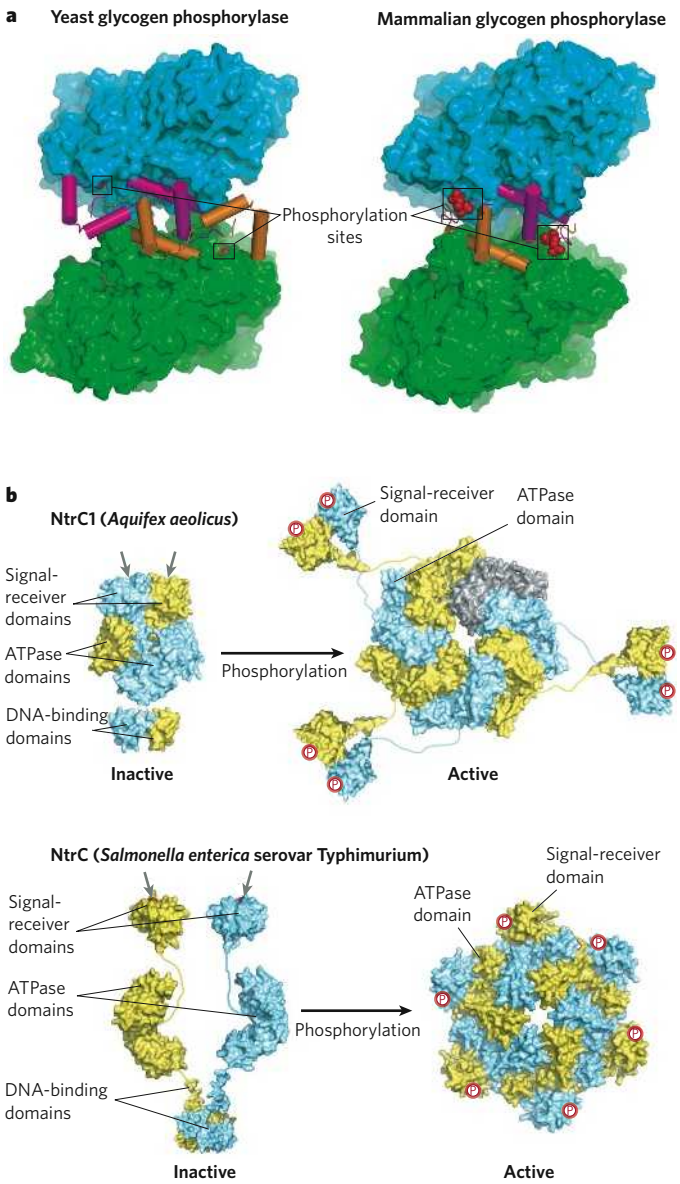


Figure 5 | Mechanisms of control by phosphorylation. a, Glycogen phosphorylase. The structures of the phosphorylated yeast enzyme (left) and mammalian enzyme (right) are shown, with one subunit in green and yellow and the other in blue and pink. Helices and loops in the N-terminal segments are shown as cylinders and coils, respectively. The phosphate groups are shown as red spheres (partly occluded in the yeast structure). This comparison shows that the structure at the N terminus (yellow and pink), which contains the sites of phosphorylation, differs between these proteins. Images generated from files from the PDB, based on data from the following: ref. 70, file 1YGP (left); and ref. 45, file 1GPA (right). **b**, Bacterial transcriptional activators of the AAA+ ATPase superfamily that associate with the σ^{54} form of RNA polymerase. Assembly of these proteins is controlled by phosphorylation, which results in a switch from the inactive (monomeric or dimeric) form (left) to the active form (right), in which the ATPase domains form an oligomeric ring. Two examples are shown, with one subunit in blue and the other in yellow. In NtrC1 from *Aquifex aeolicus* (upper panel), the signal-receiver domains hold the ATPase domains as dimers (which is an inactive conformation; left) until they are phosphorylated. After phosphorylation, a conformational change in the signal-receiver-domain dimer suppresses interaction with the ATPase domains, which are then free to assemble into the active oligomer (right). The crystal structure shows a heptamer rather than a hexamer, and the extra subunit is shown in grey. By contrast, in the homologue NtrC from *Salmonella enterica* serovar Typhimurium (lower panel), the signal-receiver domains and the ATPase domains do not interact in the absence of phosphorylation (left). After phosphorylation, the signal-receiver domains bind to a neighbouring ATPase domain, stabilizing the assembled active ring of the ATPase (right). The sites that undergo phosphorylation are indicated by grey arrows for the inactive molecules (left) and red circles for the active molecules (right). The DNA-binding domains are present at the bottom of the inactive forms of the proteins (an orientation that is required to maintain inactive NtrC in the dimeric state), and they are underneath the assembled oligomers (and therefore not visible in these diagrams).

positively (Fig. 5b). In other family members, such as NtrC1 from *Aquifex aeolicus* and dicarboxylate transport regulator D (DctD) from *Sinorhizobium meliloti*, phosphorylation is required to disrupt a dimeric state that prevents hexamerization of the ATPase domain (Fig. 5b). In this case, the signal-receiver domain is dispensable for activity. As is the case for glycogen phosphorylase, it is clear that phosphorylation-mediated control evolved after the basic mechanism of oligomerization had been set in place.

Protein tyrosine kinases

The clustering of receptor molecules at the plasma membrane is emerging as a key feature of intracellular signal transduction. Such clustering further increases the high local concentrations of membrane or receptor-associated signalling molecules, and it promotes a diverse range of protein–protein interactions[55]. Allostery is a common attribute of these proteins, with one domain modulating the activity of another domain in the same molecule. For proteins with homologous catalytic domains, these allosteric interactions follow no common pattern. This is exemplified by the protein tyrosine kinases, enzymes that are crucial for cell–cell communication in metazoans.

Cytoplasmic (non-receptor) protein tyrosine kinases have a conserved catalytic domain (known as the kinase domain), which is fused to targeting domains (also known as regulatory domains) that bind to other proteins or to lipids[56]. The primordial function of these targeting domains was probably to localize protein tyrosine kinases to sites of signalling, but they have evolved the ability to regulate the activity of the kinase domain. Here, we consider three cytoplasmic protein tyrosine kinases: Abl (the cellular homologue of the oncogene encoded by the Abelson leukaemia virus), ZAP70 (ζ-chain associated protein kinase of 70 kDa) and FAK (focal adhesion kinase; also known as PTK2). Each of these is activated by the phosphorylation of one or two tyrosine residues that are located between the targeting domains and the kinase domain, a process that releases the targeting domains from interaction with the kinase domains. In each case, however, the targeting domains suppress the activity of the kinase domain in a different manner.

Abl has two targeting domains — a Src-homology 2 (SH2) domain and an SH3 domain — fused to the kinase domain. These targeting domains clamp onto the distal surface of both lobes of the kinase domain, suppressing activity in a similar manner to that used by Src-family kinases, which are closely related[57–59] (Fig. 6a). Phosphorylation of the linker between the SH2 domain and the kinase domain in Abl prevents the engagement of the SH3–SH2 unit with the kinase domain, thereby activating Abl[60]. By contrast, ZAP70 has a tandem SH2 unit fused to the kinase domain, and this unit inhibits catalytic activity by interacting with the hinge region of the kinase domain of ZAP70, suppressing its flexibility[61] (Fig. 6a). Phosphorylation of the linker between

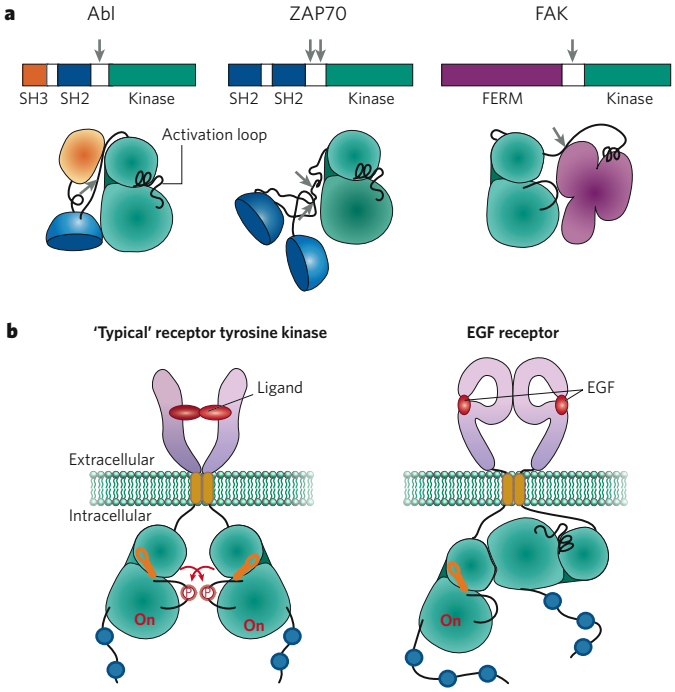


Figure 6 | Diverse regulatory mechanisms in tyrosine kinases. a, The domain organization and auto-inhibited structures of three cytoplasmic (non-receptor) protein tyrosine kinases (Abl, ZAP70 and FAK) are shown. These enzymes are activated by phosphorylation of tyrosine residues (indicated by grey arrows) located in the linker between the targeting domains and the kinase domain, but the molecular mechanism by which this phosphorylation causes activation differs for each enzyme. **b**, The activation mechanisms for a typical receptor tyrosine kinase and for the EGF receptor, an atypical receptor tyrosine kinase, are shown. Typical receptors undergo ligand-induced dimerization and can then be activated by the phosphorylation of each kinase domain by the other. For the EGF receptor, dimerization of the extracellular portion of the receptor by EGF results in the formation of an asymmetrical dimer by the kinase domains in the cytoplasm, which then activates one of these domains. This mechanism is unique to members of the EGF-receptor family. Signal transduction is propagated through the docking of SH2-containing molecules to phosphorylated residues (blue) adjacent to the kinase domains. This diagram is based on structures that were determined separately for the extracellular domains[71–73] and the cytoplasmic domain[66].

the SH2 domain and the kinase domain activates ZAP70, by preventing the formation of interactions between aromatic amino acids that are crucial for assembly of the auto-inhibited ZAP70 (ref. 61). In FAK, there is a targeting domain known as a FERM domain, which is located N-terminal to the kinase domain (Fig. 6a). Unlike the interactions in the previous two examples, the FERM domain interacts with the 'front' of the kinase domain, where it directly blocks access to the active site[62]. Phosphorylation of the linker activates FAK by destabilizing the interaction between the FERM domain and the kinase domain. This comparison of three cytoplasmic protein tyrosine kinases is representative of a feature common to signalling pathways: they resemble the haphazard collection of assorted parts in the imagined devices of the cartoonist Rube Goldberg, with nature inventing multiple ways to regulate the activity of a conserved catalytic domain.

Receptor tyrosine kinases are transmembrane proteins in which an extracellular ligand-binding domain is separated from an intracellular kinase domain by the plasma membrane[63]. The simplest mechanism (most probably the primordial mechanism) by which a receptor tyrosine kinase can be activated involves the phosphorylation of a centrally located activation loop in one subunit of a homodimer by the kinase domain in the other subunit, and vice versa, a reaction that is promoted by ligand-induced dimerization[64]. Individual receptors, however, have evolved regulatory mechanisms that are layered on top of this simple mechanism. The insulin receptor, for example, is a covalently crosslinked dimer, and *trans*-phosphorylation results from an insulin-induced conformational change rather than from a monomer–dimer transition[65]. Unlike the insulin receptor and other typical receptor tyrosine kinases (Fig. 6b), the receptor for epidermal growth factor (EGF) does not require phosphorylation of an activation loop for catalytic activity. Instead, the activation mechanism involves an asymmetrical interaction between the large lobe of one kinase domain and the small lobe of the other, a process that stabilizes the active conformation of the latter[66] (Fig. 6b). This mechanism, which resembles the way in which protein kinases that control the cell cycle (CDKs) are activated by cyclins, does not seem to be used by other receptor tyrosine kinases. Within the EGF-receptor family, however, the ability of the kinase domains to function as both activators and transducers for each other leads to a powerful combinatorial response to a variety of ligands.

Conclusions

Genome sequencing has only begun to uncover the molecular details of the great puzzle of how complex and interacting molecular forms emerged from simpler ones. One of the findings that has emerged from genomic analysis is that the machinery of life is conserved across the evolutionary tree. Globin subunits, for example, have the same overall structure and the same chemical linkage to the haem iron in plants, invertebrates and mammals. Glycogen phosphorylase has the same dimeric structure in yeast and humans. Beginning with haemoglobin, researchers have come to appreciate that the regulated functioning of protein-based machines depends on allosteric interactions between one or more components in the assembly. Given the uniformity of the basic designs of protein modules, it could be expected that the allosteric mechanisms are also conserved. That they are not helps to resolve the otherwise insurmountable paradox of how such intricate mechanisms could have evolved from the constituent parts. The physical imperative for the allosteric control of oxygen binding to its transport protein has been solved by evolution in many organisms, but there is no combinatorial imperative that requires a particular interface or residue to be used in the mechanism. The variety of mechanisms that has been found seems to disclose the random nature of the events that gave rise to each.

Whether this 'rule of varied allosteric control' is generally applicable should emerge from further comparative studies of allosteric control in protein families. Similarly, genome sequencing of the organisms that diverged earliest might reveal ancestors of present-day proteins that gave rise to protein interactions. In particular, such analyses might definitively determine whether protein–protein interactions arise from the fusion of genes encoding protein domains followed by the fission of such fused genes. This hypothesis could be termed the 'rule of heterodimer evolution by protein fission'. It is now the turn of molecular scientists to uncover details of the process that Charles Darwin summarized famously in the final sentence of *On the Origin of Species*: "whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being evolved". ■

1. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402,** C47–C52 (1999).
2. Monod, J., Changeux, J. P. & Jacob, F. Allosteric proteins and cellular control systems. *J. Mol. Biol.* **6,** 306–329 (1963).
3. Beernink, P. T., Endrizzi, J. A., Alber, T. & Schachman, H. K. Assessment of the allosteric mechanism of aspartate transcarbamoylase based on the crystalline structure of the unregulated catalytic subunit. *Proc. Natl Acad. Sci. USA* **96,** 5388–5393 (1999).
4. Behe, M. J. *Darwin's Black Box: The Biochemical Challenge to Evolution* (The Free Press, New York, 2003).
5. Perutz, M. F. Stereochemistry of cooperative effects in haemoglobin. *Nature* **228,** 726–739 (1970).
6. Xu, D., Tsai, C. J. & Nussinov, R. Mechanism and evolution of protein dimerization. *Protein Sci.* **7,** 533–544 (1998).
7. Ispolatov, I., Yuryev, A., Mazo, I. & Maslov, S. Binding properties and evolution of homodimers in protein–protein interaction networks. *Nucleic Acids Res.* **33,** 3629–3635 (2005).
8. Liang, J., Kim, J. R., Boock, J. T., Mansell, T. J. & Ostermeier, M. Ligand binding and allostery can emerge simultaneously. *Protein Sci.* **16,** 929–937 (2007).
9. Pawson, T. & Scott, J. D. Signaling through scaffold, anchoring, and adaptor proteins. *Science* **278,** 2075–2080 (1997).

10. Klemm, J. D. & Pabo, C. O. Oct-1 POU domain–DNA interactions: cooperative binding of isolated subdomains and effects of covalent linkage. *Genes Dev.* **10,** 27–36 (1996).
11. Robinson, C. R. & Sauer, R. T. Covalent attachment of Arc repressor subunits by a peptide linker enhances affinity for operator DNA. *Biochemistry* **35,** 109–116 (1996).
12. Predki, P. F. & Regan, L. Redesigning the topology of a four-helix-bundle protein: monomeric Rop. *Biochemistry* **34,** 9834–9839 (1995).
13. Liang, H., Sandberg, W. S. & Terwilliger, T. C. Genetic fusion of subunits of a dimeric protein substantially enhances its stability and rate of folding. *Proc. Natl Acad. Sci. USA* **90,** 7010–7014 (1993).
14. Pedersen, S., Bloch, P. L., Reeh, S. & Neidhardt, F. C. Patterns of protein synthesis in *E. coli*: a catalog of the amount of 140 individual proteins at different growth rates. *Cell* **14,** 179–190 (1978).
15. Lu, P., Vogel, C., Wang, R., Yao, X. & Marcotte, E. M. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nature Biotechnol.* **25,** 117–124 (2007).
16. Noguchi, C. T. & Schechter, A. N. Sickle hemoglobin polymerization in solution and in cells. *Annu. Rev. Biophys. Biophys. Chem.* **14,** 239–263 (1985).
17. Fermi, G., Perutz, M. F., Shaanan, B. & Fourme, R. The crystal structure of human deoxyhaemoglobin at 1.74 Å resolution. *J. Mol. Biol.* **175,** 159–174 (1984).
18. Bahadur, R. P., Chakrabarti, P., Rodier, F. & Janin, J. Dissecting subunit interfaces in homodimeric proteins. *Proteins* **53,** 708–719 (2003).
19. Fersht, A. R. *Structure and Mechanism in Protein Science* (Freeman, New York, 1999).
20. Eisenberg, D., Wesson, M. & Yanashita, M. Interpretation of protein folding and binding with atomic solvation parameters. *Chem. Scr.* **29A,** 217–221 (1989).
21. Behe, M. J. *The Edge of Evolution* (Free Press, New York, 2007).
22. Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature* **402,** 83–86 (1999).
23. Bennett, M. J., Choe, S. & Eisenberg, D. Domain swapping: entangling alliances between proteins. *Proc. Natl Acad. Sci. USA* **91,** 3127–3131 (1994).
24. Schlunegger, M. P., Bennett, M. J. & Eisenberg, D. Oligomer formation by 3D domain swapping: a model for protein assembly and misassembly. *Adv. Protein Chem.* **50,** 61–122 (1997).
25. Finn, R. D. *et al.* Pfam: clans, web tools and services. *Nucleic Acids Res.* **34,** D247–D251 (2006).
26. Servant, F. *et al.* ProDom: automated clustering of homologous domains. *Brief. Bioinform.* **3,** 246–251 (2002).
27. Enright, A. J., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402,** 86–90 (1999).
28. Thoden, J. B., Raushel, F. M., Benning, M. M., Rayment, I. & Holden, H. M. The structure of carbamoyl phosphate synthetase determined to 2.1 Å resolution. *Acta Crystallogr. D Biol. Crystallogr.* **55,** 8–24 (1999).
29. Snel, B., Bork, P. & Huynen, M. Genome evolution: gene fusion versus gene fission. *Trends Genet.* **16,** 9–11 (2000).
30. Kummerfeld, S. K. & Teichmann, S. A. Relative rates of gene fusion and fission in multi-domain proteins. *Trends Genet.* **21,** 25–30 (2005).
31. Fong, J. H., Geer, L. Y., Panchenko, A. R. & Bryant, S. H. Modeling the evolution of protein domain architectures using maximum parsimony. *J. Mol. Biol.* **366,** 307–315 (2007).
32. Pabo, C. O. & Sauer, R. T. Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.* **61,** 1053–1095 (1992).
33. Wolberger, C. Multiprotein–DNA complexes in transcriptional regulation. *Annu. Rev. Biophys. Biomol. Struct.* **28,** 29–56 (1999).
34. Panne, D., Maniatis, T. & Harrison, S. C. An atomic model of the interferon-β enhanceosome. *Cell* **129,** 1111–1123 (2007).
35. Wilson, D. S., Guenther, B., Desplan, C. & Kuriyan, J. High resolution crystal structure of a paired (Pax) class cooperative homeodomain dimer on DNA. *Cell* **82,** 709–719 (1995).
36. Passner, J. M., Ryoo, H. D., Shen, L., Mann, R. S. & Aggarwal, A. K. Structure of a DNA-bound Ultrabithorax–Extradenticle homeodomain complex. *Nature* **397,** 714–719 (1999).
37. Piper, D. E., Batchelor, A. H., Chang, C. P., Cleary, M. L. & Wolberger, C. Structure of a HoxB1–Pbx1 heterodimer bound to DNA: role of the hexapeptide and a fourth homeodomain helix in complex formation. *Cell* **96,** 587–597 (1999).
38. LaRonde-LeBlanc, N. A. & Wolberger, C. Structure of HoxA9 and Pbx1 bound to DNA: Hox hexapeptide and DNA recognition anterior to posterior. *Genes Dev.* **17,** 2060–2072 (2003).
39. Klemm, J. D., Rould, M. A., Aurora, R., Herr, W. & Pabo, C. O. Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA-binding modules. *Cell* **77,** 21–32 (1994).
40. Jacobson, E. M., Li, P., Leon-del-Rio, A., Rosenfeld, M. G. & Aggarwal, A. K. Structure of Pit-1 POU domain bound to DNA as a dimer: unexpected arrangement and flexibility. *Genes Dev.* **11,** 198–212 (1997).
41. Li, T., Stark, M. R., Johnson, A. D. & Wolberger, C. Crystal structure of the MAT**a**1/MATα2 homeodomain heterodimer bound to DNA. *Science* **270,** 262–269 (1995).
42. Tan, S. & Richmond, T. J. Crystal structure of the yeast MATα2/MCM1/DNA ternary complex. *Nature* **391,** 660–666 (1998).
43. Royer, W. E., Zhu, H., Gorr, T. A., Flores, J. F. & Knapp, J. E. Allosteric hemoglobin assembly: diversity and similarity. *J. Biol. Chem.* **280,** 27477–27480 (2005).
44. Royer, W. E., Hendrickson, W. A. & Chiancone, E. The 2.4 Å crystal structure of *Scapharca* dimeric hemoglobin. *J. Biol. Chem.* **264,** 21052–21061 (1989).
45. Barford, D., Hu, S. H. & Johnson, L. N. Structural mechanism for glycogen phosphorylase control by phosphorylation and AMP. *J. Mol. Biol.* **218,** 233–260 (1991).
46. Barford, D. & Johnson, L. N. The allosteric transition of glycogen phosphorylase. *Nature* **340,** 609–616 (1989).
47. Sprang, S. R. *et al.* Structural changes in glycogen phosphorylase induced by phosphorylation. *Nature* **336,** 215–221 (1988).
48. Buchbinder, J. L., Rath, V. L. & Fletterick, R. J. Structural relationships among regulated and unregulated phosphorylases. *Annu. Rev. Biophys. Biomol. Struct.* **30,** 191–209 (2001).
49. Rath, V. L. & Fletterick, R. J. Parallel evolution in two homologues of phosphorylase. *Nature Struct. Biol.* **1,** 681–690 (1994).
50. Palm, D., Goerl, R. & Burger, K. J. Evolution of catalytic and regulatory sites in phosphorylases. *Nature* **313,** 500–502 (1985).
51. Hwang, P. K. & Fletterick, R. J. Convergent and divergent evolution of regulatory sites in eukaryotic phosphorylases. *Nature* **324,** 80–84 (1986).
52. Lee, S. Y. *et al.* Regulation of the transcriptional activator NtrC1: structural studies of the regulatory and AAA+ ATPase domains. *Genes Dev.* **17,** 2552–2563 (2003).
53. Doucleff, M. *et al.* Negative regulation of AAA+ ATPase assembly by two component receiver domains: a transcription activation mechanism that is conserved in mesophilic and extremely hyperthermophilic bacteria. *J. Mol. Biol.* **353,** 242–255 (2005).
54. De Carlo, S. *et al.* The structural basis for regulated assembly and function of the transcriptional activator NtrC. *Genes Dev.* **20,** 1485–1495 (2006).
55. Pawson, T. & Nash, P. Assembly of cell regulatory systems through protein interaction domains. *Science* **300,** 445–452 (2003).
56. Neet, K. & Hunter, T. Vertebrate non-receptor protein-tyrosine kinase families. *Genes Cells* **1,** 147–169 (1996).
57. Nagar, B. *et al.* Structural basis for the autoinhibition of c-Abl tyrosine kinase. *Cell* **112,** 859–871 (2003).
58. Sicheri, F., Moarefi, I. & Kuriyan, J. Crystal structure of the Src-family tyrosine kinase Hck. *Nature* **385,** 602–609 (1997).
59. Xu, W., Harrison, S. C. & Eck, M. J. Three-dimensional structure of the tyrosine kinase c-Src. *Nature* **385,** 595–602 (1997).
60. Brasher, B. B. & Van Etten, R. A. c-Abl has high intrinsic tyrosine kinase activity that is stimulated by mutation of the Src homology 3 domain and by autophosphorylation at two distinct regulatory sites. *J. Biol. Chem.* **275,** 35631–35637 (2000).
61. Deindl, S. *et al.* Structural basis for the inhibition of tyrosine kinase activity of ZAP-70. *Cell* **129,** 735–746 (2007).
62. Lietha, D. *et al.* Structural basis for the autoinhibition of focal adhesion kinase. *Cell* **129,** 1177–1187 (2007).
63. Schlessinger, J. Cell signaling by receptor tyrosine kinases. *Cell* **103,** 211–225 (2000).
64. Schlessinger, J. Ligand-induced, receptor-mediated dimerization and activation of EGF receptor. *Cell* **110,** 669–672 (2002).
65. Hubbard, S. R. & Till, J. H. Protein tyrosine kinase structure and function. *Annu. Rev. Biochem.* **69,** 373–398 (2000).
66. Zhang, X., Gureasko, J., Shen, K., Cole, P. A. & Kuriyan, J. An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor. *Cell* **125,** 1137–1149 (2006).
67. Deeds, E. J., Ashenberg, O., Gerardin, J. & Shakhnovich, E. I. Robust protein–protein interactions in crowded cellular environments. *Proc. Natl Acad. Sci. USA* **104,** 14952–14957 (2007).
68. Royer, W. E.. High-resolution crystallographic analysis of a co-operative dimeric hemoglobin. *J. Mol. Biol.* **235,** 657–681 (1994).
69. Flores, J. F. *et al.* Sulfide binding is mediated by zinc ions discovered in the crystal structure of a hydrothermal vent tubeworm hemoglobin. *Proc. Natl Acad. Sci. USA* **102,** 2713–2718 (2005).
70. Lin, K., Rath, V. L., Dai, S. C., Fletterick, R. J. & Hwang, P. K. A protein phosphorylation switch at the conserved allosteric site in GP. *Science* **273,** 1539–1542 (1996).
71. Cho, H. S. & Leahy, D. J. Structure of the extracellular region of HER3 reveals an interdomain tether. *Science* **297,** 1330–1333 (2002).
72. Garrett, T. P. *et al.* Crystal structure of a truncated epidermal growth factor receptor extracellular domain bound to transforming growth factor α. *Cell* **110,** 763–773 (2002).
73. Ogiso, H. *et al.* Crystal structure of the complex of human epidermal growth factor and receptor extracellular domains. *Cell* **110,** 775–787 (2002).

Acknowledgements We are grateful to A. K. Aggarwal, R. E. Dickerson, S. C. Harrison, L. N. Johnson, E. M. Marcotte, M. Robertson, W. E. Royer, M. Seeliger, D. E. Wemmer, C. Wolberger, T. O. Yeates and many other colleagues for comments. We thank S. Deindl, L. Leighton, W. E. Royer, D. E, Wemmer and X. Zhang for assistance with the figures. Support from the Howard Hughes Medical Institute, the National Institutes of Health, the US Department of Energy Office of Biological & Environmental Research, and the National Science Foundation is gratefully acknowledged.

Author information Reprints and permissions information is available at npg.nature.com/reprints. Correspondence should be addressed to the authors (david@mbi.ucla.edu; kuriyan@berkeley.edu).

The biological impact of mass-spectrometry-based proteomics

Benjamin F. Cravatt[1,2], Gabriel M. Simon[1,2] & John R. Yates III[1]

In the past decade, there have been remarkable advances in proteomic technologies. Mass spectrometry has emerged as the preferred method for in-depth characterization of the protein components of biological systems. Using mass spectrometry, key insights into the composition, regulation and function of molecular complexes and pathways have been gained. From these studies, it is clear that mass-spectrometry-based proteomics is now a powerful 'hypothesis-generating engine' that, when combined with complementary molecular, cellular and pharmacological techniques, provides a framework for translating large data sets into an understanding of complex biological processes.

Of the many fascinating discoveries made by genome-sequencing projects, perhaps none is more provocative than the prediction that all prokaryotes and eukaryotes produce numerous proteins with uncharacterized or pleiotropic functions[1,2]. Confronted with the challenge of annotating this enormous segment of the proteome, scientists have sought to expedite the characterization of proteins by developing new methods for rapid and parallel analysis. These large-scale approaches to protein science are collectively termed proteomics[3,4].

Proteomics encompasses diverse techniques that allow different aspects of protein structure and function to be analysed. Many proteomic methods — including protein microarrays[5,6], large-scale two-hybrid analyses[7], and high-throughput protein production and crystallization[8] — have had a marked impact on the current understanding of protein structures, activities and interactions. Among proteomic techniques, however, mass spectrometry has emerged as the main method for analysing the production and function of proteins in native biological systems[9–11].

Mass spectrometry has become the dominant technique for several reasons, mainly because of its unparalleled ability to acquire high-content quantitative information about biological samples of enormous complexity. The core technologies of mass-spectrometry-based proteomics, including the instrumentation and the methods for data acquisition and analysis, have been discussed in several recent reviews[9–11] and are outlined in Box 1. Although these technologies will continue to be developed in the quest for improved sensitivity, throughput and proteome coverage, mass-spectrometry-based proteomics has now developed to the point at which it is routinely applied worldwide to address a large range of biological problems. It therefore seems an opportune time to reflect on the functional impact of mass-spectrometry-based proteomics. What has been learned about the molecular mechanisms of complex biological processes? How were successful experiments carried out? What additional methods were required to make important biological discoveries? Finally, are there lessons that might guide future applications?

In this review, we address these questions by focusing on several cases in which mass-spectrometry-based proteomics has had a crucial role in advancing our understanding of basic cellular and physiological processes. We highlight common themes that seem to have primed investigations for success, including the configuration of biologically relevant model systems (and controls), the implementation of mass-spectrometry-based proteomics as a hypothesis-generating platform, and a commitment to focused follow-up studies that test emerging hypotheses directly. These examples underscore both the opportunities and the challenges that face the systematic integration of mass-spectrometry-based proteomic techniques into the arsenal of experimental approaches used by molecular and cellular biologists.

Mass-spectrometry-based proteomics has several biological applications. In many pioneering studies, it was used to make an inventory of the content of subcellular structures and organelles, creating valuable repositories of information about the localization of proteins in cells and tissues (see ref. 12 for a recent review). It is also emerging as a powerful way to discern higher-order structural features of protein complexes, including subunit orientation and stoichiometry (see page 973). Here, we focus on two main applications — the functional characterization of protein complexes, and the functional characterization of protein pathways — highlighting studies that have led to major advances in understanding the molecular basis of cellular and physiological processes.

Functional characterization of protein complexes

Many proteins function as components of complexes in cells and tissues[13]. Protein complexes can vary in size and composition, from megadalton assemblies of dozens of proteins (such as the ribosome and the spliceosome) to smaller clusters of a few proteins. The composition and stability of protein complexes is highly regulated in both a context-dependent manner (for example, there are cell-type-specific differences) and a time-dependent manner[14]. These biological variables present a challenge to researchers interested in determining the structure and the function of protein complexes. Mass-spectrometry-based proteomics, however, can address this issue in a systematic and relatively unbiased manner, often revealing surprising protein partnerships and assemblies that regulate cellular and physiological processes. In addition to the examples discussed in this section, other notable studies that have used mass-spectrometry-based proteomics to characterize protein complexes are described in refs 15–19.

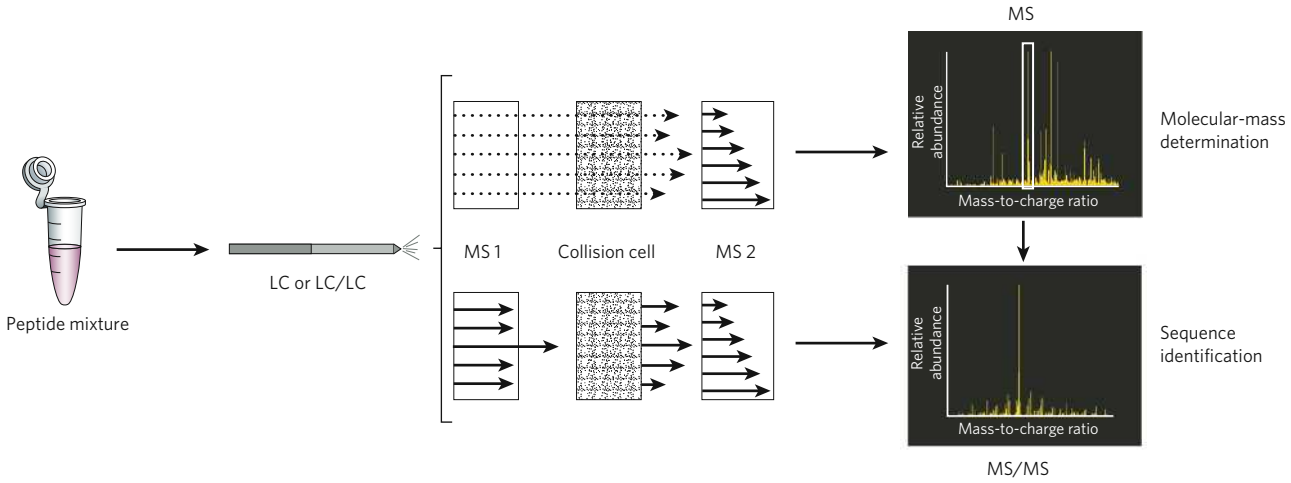
A mitochondrial protein complex that links apoptosis and glycolysis

Stanley Korsmeyer and colleagues[20] provided an early example of the value of mass-spectrometry-based proteomics for uncovering unexpected

[1]Department of Chemical Physiology, [2]The Skaggs Institute for Chemical Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA.

physical connections between proteins that had been thought to function in independent pathways. With the aim of mapping mitochondrial protein complexes that contain the pro-apoptotic protein Bcl-2 antagonist of cell death (BAD), they developed a gentle and efficient enrichment method to isolate these complexes from mouse liver mitochondria (Fig. 1a). A 232-kDa assembly consisting of five major proteins was identified by silver staining of proteins separated by polyacrylamide gel electrophoresis (PAGE). The protein spots constituting this complex were then excised and analysed by liquid-chromatography–tandem mass spectrometry (LC–MS/MS). This revealed, in addition to BAD, the presence of protein phosphatase 1C (PP1C), cyclic-AMP-dependent protein kinase (PKA) and the PKA-anchoring protein WAVE1. Previous studies had implicated PKA as an important regulator of BAD, inhibiting the activity of BAD by phosphorylating multiple serine residues[21]. The authors provided

Box 1 | Fundamentals of mass-spectrometry-based proteomic experiments



Mass-spectrometry-based proteomic experiments involve several steps (see Figure). A peptide mixture can be obtained from the sample of interest by proteolytic digestion of a protein mixture or of a gel band or spot, following separation by electrophoresis. These peptides are then introduced into a one-dimensional (LC) or multi-dimensional (LC/LC) liquid-chromatography system. After separation, they are eluted into an electrospray ionization tandem mass spectrometer. The mass-to-charge ratios of the peptide ions are measured first by mass spectrometry (upper panels; ions pass unperturbed through the first mass analyser and collision cell) to determine the molecular mass of each peptide. Then, each peptide ion is isolated in the first mass analyser (MS 1) and directed into a collision cell, where it collides with neutral gas molecules (for example, helium) and becomes fragmented (lower panels). The mass-to-charge ratios of the resultant fragments are measured in the second mass analyser (MS 2), producing a tandem mass spectrum (shown for the peptide ion indicated with a white rectangle in the upper panel) and, after computer analysis, an amino-acid sequence for each peptide. These steps are described in further detail below.

Sample preparation

Two general strategies are often used to prepare proteins. Proteins that have been enriched or obtained as part of an experiment can be fractionated by SDS–polyacrylamide gel electrophoresis (SDS–PAGE). Individual bands can be removed and analysed, or an entire lane can be excised and divided into 10–15 slices. Proteins in the gel slices are then digested *in situ* with trypsin, and the peptides are extracted. Extracted peptides are then analysed by mass spectrometry. Protein mixtures can also be digested directly in solution. A protein mixture is denatured by using chaotropes and then digested — sometimes by a two-step procedure that involves proteases, such as the endoprotease LysC followed by trypsin — to generate a peptide mixture that is suitable for mass-spectrometry analysis. In general, trypsin digestion is preferred to generate peptides with an arginine or lysine residue at the C terminus, but other types of enzyme, including nonspecific proteases, have also been used[78].

Liquid chromatography

Before peptide mixtures are introduced into the mass spectrometer, they are fractionated in-line with the instrument. The most common method of fractionation is reversed-phase liquid chromatography, which separates peptides according to their hydrophobicity. To achieve the best sensitivity and efficiency of separation, microcolumns (<100 μm in length) with a small diameter tip (for example, 5 μm) are typically used. The electrospray ionization that takes place in the mass spectrometer acts like a concentration-dependent detector; therefore, the introduction of peptides in narrow peaks improves detection limits, and low flow rates (in the order of nL min^{-1}) are used to achieve this. As peptide mixtures become more complex, the introduction of a second dimension of separation can improve the resolution of separation. A good choice for a second dimension is strong cation-exchange (SCX) liquid chromatography, which separates peptides mainly on the basis of positive charges. SCX can be used off-line, and then each fraction is analysed by reversed-phase high-pressure liquid chromatography, followed by mass spectrometry. Alternatively, both the reversed-phase and SCX resins can be packed into a single column, and, by introducing buffers in series, a two-dimensional separation can be achieved before mass-spectrometry analysis.

Electrospray ionization

A potential is placed on the liquid flowing from the liquid-chromatography column through a fused silica column or needle, causing the solution to spray. The spray contains fine droplets that encompass the sample. The droplets are desolvated as they enter the mass spectrometer, by applying heat to generate ions. The efficiency of ionization depends on the chemical properties of each molecule.

Mass-spectrometry analysis

Mass spectrometers measure the mass-to-charge ratio of an ion. This is carried out by manipulating ions in electric and/or magnetic fields or by measuring their time of flight (TOF). In addition to determining the mass-to-charge ratio, the intensity of the signal obtained reflects the abundance of the ion. The abundance of ions can vary with the ionization, so samples can be labelled with stable isotopes to determine quantitatively the ratio of peptides from different 'states' (by measuring the mass-to-charge ratio and abundance). Various mass analysers are used in proteomic experiments, including ion-trap mass spectrometers, quadrupole/TOF hybrids, ion-trap/orbitrap hybrids and ion-trap/ion-cyclotron-resonance (FTMS) hybrids. Some types of mass analyser can measure the mass-to-charge ratio with high resolution (up to 150,000 $m/\Delta m$, where m denotes mass) and high mass accuracy (to <1 part per million).

strong evidence that PP1C functions to counter the effect of PKA, by dephosphorylating BAD. These data therefore led to a model in which a BAD–PKA–PP1C complex, possibly scaffolded by WAVE1, creates a local microenvironment in which the phosphorylation status of BAD can be finely controlled.

Korsmeyer and colleagues next considered the possible function(s) of this BAD-containing complex in mitochondrial physiology. First, they pursued the characterization of the fifth (and still unidentified) component of the complex. By LC–MS/MS analysis, this 50-kDa protein was identified as the glycolytic enzyme glucokinase (also known as hexokinase 4). This result was initially surprising; even though apoptosis and glycolysis are both crucial physiological processes that are linked to cell survival[22,23], the molecular pathways involved had been thought to function independently. The organization of glucokinase and BAD into a stable multiprotein complex in mitochondria indicated otherwise. Indeed, the authors showed that *Bad*$^{-/-}$ mice, which lack the gene encoding BAD, had severely blunted mitochondrial glucokinase activity, glucose-driven respiration and glucose-dependent ATP production. Moreover, these mice had significantly higher blood glucose concentrations after fasting than did wild-type (*Bad*$^{+/+}$)mice. The effect of BAD on glucose homeostasis depended on its phosphorylation status, suggesting that other members of the BAD-containing complex (for example, PKA and PP1C) had a regulatory role.

These findings therefore indicate that the mitochondrial fraction of glucokinase — defined as that portion of the enzyme stably associated with the BAD–PKA–PP1C–WAVE1 complex (Fig. 1b) — has a disproportionate role in maintaining proper glucose metabolism (given that most of the glucokinase in a cell is cytosolic). The glucokinase- and BAD-containing mitochondrial complex was also proposed to function as an integration centre that links metabolic state and cell death. This hypothesis was supported by the finding that liver cells from *Bad*$^{-/-}$ mice underwent less glucose-deprivation-induced apoptosis than wild-type liver cells. In summary, the mass-spectrometry-based proteomic analysis of mitochondrial BAD-containing complexes discovered an unexpected physical association between a key apoptotic protein (BAD) and a key glycolytic protein (glucokinase), thereby leading to new models to explain how cells coordinate metabolic signals and survival signals.

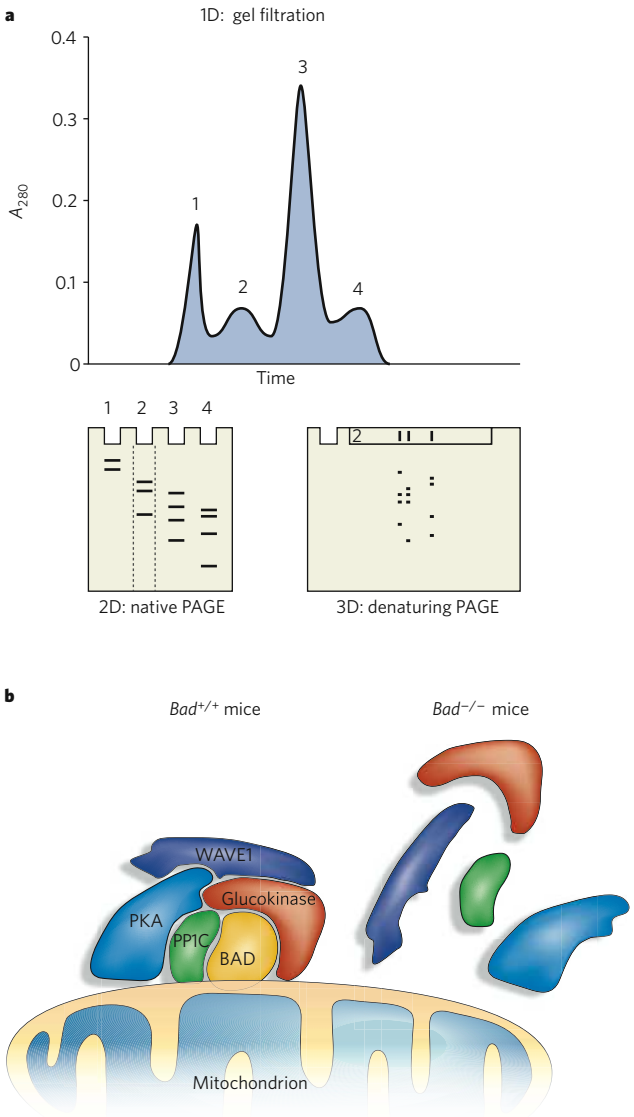


Figure 1 | Discovery of a BAD-containing protein complex that localizes to mitochondria and integrates apoptotic and glycolytic processes. **a**, A complex containing BAD, PKA, PP1C, WAVE1 and glucokinase was discovered by multidimensional fractionation of mitochondrial protein complexes from *Bad*$^{+/+}$ (wild-type) mice and *Bad*$^{-/-}$ mice, followed by LC–MS/MS analysis[20]. Multidimensional fractionation involved gel filtration as the first dimension (1D), native (non-denaturing) PAGE as the second dimension (2D), and denaturing PAGE (SDS–PAGE) as the third dimension (3D). **b**, In the absence of BAD, none of the members of the protein complex associates with mitochondrial membranes.

A transcription-factor complex relevant to trichothiodystrophy

The proper maintenance, repair and transcription of DNA requires several multiprotein complexes[24]. Ruedi Aebersold and colleagues[25] were interested in fully characterizing the components of the yeast (*Saccharomyces cerevisiae*) RNA polymerase II (PolII) pre-initiation complex and established an elegant biochemical system to enrich this protein assembly (Fig. 2). They first isolated nuclear extracts from a mutant yeast strain carrying a temperature-sensitive allele of the TATA-binding protein (TBP) and incubated these proteomes with an immobilized *HIS4* promoter from yeast, which includes the TATA box, in the presence or absence of recombinant TBP. They then used isotope-coded affinity tagging (ICAT) in conjunction with LC–MS/MS[26] to identify proteins that were quantitatively enriched (by at least 1.9-fold) in promoter-associated fractions from TBP-containing proteomes. Nearly all of the proteins that met this criterion were known components of the PolII transcriptional machinery, with the notable exception of an open reading frame — YDR079C-A (Fig. 2) — which corresponded to a small (8 kDa) protein of unknown function.

BLAST searches (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi) revealed homologues of YDR079C-A in several eukaryotic organisms, including humans and *Chlamydomonas reinhardtii*. Interestingly, the protein encoded by *C. reinhardtii* was known to suppress sensitivity to ultraviolet light (which can damage DNA)[27]. This finding suggested that the protein encoded by YDR079C-A might be a component of transcription factor IIH (TFIIH), which has a role in both general transcription and repair of DNA damage. Aebersold and colleagues[25] confirmed this hypothesis by carrying out a further round of quantitative proteomic experiments, this time comparing proteins that bound to an epitope (FLAG)-tagged YDR079C-A protein. The proteins that were most enriched after immunoprecipitation of FLAG–YDR079C-A protein with FLAG-specific antibodies were components of TFIIH. Conversely, immunoprecipitation with antibodies specific for other TFIIH components 'pulled down' the YDR079C-A protein. Finally, YDR079C-A protein was shown to be required for stable recruitment of TFIIH to promoters. These findings confirmed that the YDR079C-A protein is a core subunit of TFIIH, prompting the authors to rename the protein as the transcription factor subunit Tfb5.

In an amazing example of a basic scientific discovery being rapidly translated, the human homologue of Tfb5 was almost immediately shown to be the long-sought gene product that is mutated in a set of unexplained cases of trichothiodystrophy[28], a rare human disease characterized by brittle hair and skin photosensitivity. Most cases of trichothiodystrophy had been traced to mutations in the genes encoding the nine known components of TFIIH. However, an individual with symptoms of trichothiodystrophy but no mutations in these genes had been found several years earlier[29]. Interestingly, cells from patients

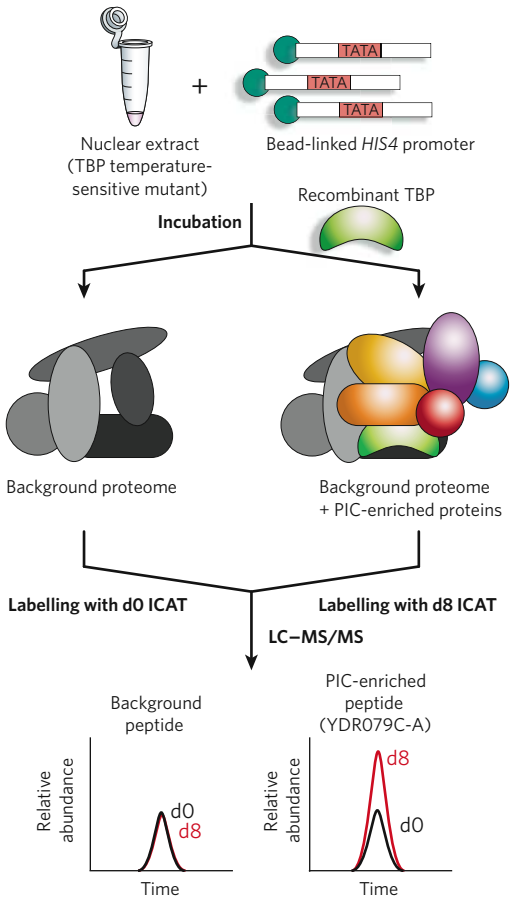


Figure 2 | Discovery of Tfb5 as the tenth subunit of TFIIH, which is involved in transcriptional and DNA-repair processes. Nuclear extracts from yeast expressing a temperature-sensitive mutant of TBP were incubated with *HIS4*-promoter-linked beads, in the presence or absence of recombinant TBP. The inclusion of recombinant TBP facilitated enrichment for proteins that are members of PolII pre-initiation complexes (PICs). Samples enriched in the absence or presence of recombinant TBP were then labelled with non-deuterated (d0) and deuterated (d8) ICAT probes, respectively, enabling LC–MS-based quantitative proteomic analysis of differentially enriched proteins. This led to the identification of YDR079C-A (subsequently named Tfb5) as a component of the transcription-factor complex TFIIH[25]. Complementary genetic studies (not shown) confirmed that the gene encoding the human homologue of Tfb5 is mutated in rare forms of trichothiodystrophy[28].

with this unusual variant of trichothiodystrophy (called trichothiodystrophy A) were found to have low cellular concentrations of TFIIH[29]. On discovery of Tfb5 as the tenth component of yeast TFIIH, Wim Vermeulen, Aebersold and colleagues sequenced the corresponding human gene in patients with trichothiodystrophy A and found inactivating mutations in four individuals from three separate families[28]. Moreover, they showed that recombinant human TFB5 could stabilize TFIIH complexes and correct the DNA-repair defects in cells from patients with trichothiodystrophy A.

In summary, the use of mass-spectrometry-based proteomics to characterize a previously unknown component of TFIIH catalysed a remarkable bench-to-bedside-to-bench investigation that succeeded in explaining the molecular basis for a human photosensitivity syndrome. On a technical note, it is worth emphasizing that this discovery hinged on the use of quantitative profiling, which allowed the researchers to confidently identify Tfb5, despite its showing only a moderate (twofold) increase in abundance in promoter-associated samples compared with control samples. Direct LC–MS/MS analysis also proved crucial, because the small size of Tfb5 (8 kDa) precluded straightforward detection by SDS–PAGE (and might explain why this protein eluded detection by more-classical methods). Finally, the interaction of Tfb5 with another component of TFIIH, Tfb2, was recently confirmed in a genome-wide tandem-affinity-purification study[30], thus underscoring the capacity of large-scale mass-spectrometry-based proteomic experiments to characterize physiologically relevant protein complexes.

A chaperone complex that regulates CFTR folding and transport

Transmembrane proteins depend on a complex range of chaperones and co-chaperones for optimal folding, localization and, ultimately, function[31,32]. Genetic mutations that impair the folding of membrane proteins form the basis of many human diseases, including cystic fibrosis. Cystic fibrosis is mainly caused by point mutations in the gene encoding an apical membrane ATP-regulated chloride channel, which is known as the cystic fibrosis transmembrane conductance regulator (CFTR)[33]. The main disease-associated mutation, ΔF508 (deletion of the phenylalanine residue at position 508 of the wild-type protein), disrupts the folding of CFTR in the endoplasmic reticulum, leading to almost complete degradation of this channel[34] (Fig. 3a). Interestingly, however, properly folded CFTR with this mutation can traffic to the plasma membrane, where it forms a functional chloride channel. These findings suggest that rescuing the folding of ΔF508-CFTR could eventually be used to treat patients with cystic fibrosis.

Initial studies have implicated both chaperone assemblies that contain heat-shock protein 90 (HSP90) and those that contain HSP40 and HSP70 in the folding pathways for CFTR[35,36]. William Balch, John Yates and colleagues[37] proposed that a more complete understanding of the chaperone assemblies that regulate CFTR folding and transport could be achieved by carrying out a proteomic analysis of the proteins associated with the channel. The authors used the shotgun LC–MS method MudPIT (multidimensional protein-identification technology)[38] to analyse cells expressing the gene encoding the wild-type CFTR or ΔF508-CFTR. MudPIT analysis of wild-type CFTR and ΔF508-CFTR immunoprecipitates identified nearly 200 CFTR-associated proteins (compared with controls in which nonspecific antibodies or cells lacking CFTR were used). Collectively, these proteins have been named the CFTR interactome (Fig. 3b). These proteins included known CFTR-binding chaperones, such as calnexin, HSP40–HSP70 and HSP90, as well as many previously unknown interactors.

These researchers next set out to test whether any of the newly discovered CFTR-associated proteins regulated channel folding and export from the endoplasmic reticulum[37]. RNA-interference (RNAi)-mediated knockdown of either p23 (also known as PTGES3) or FKBP8 — two HSP90 co-chaperones that were selectively identified in ΔF508-CFTR immunoprecipitates (as determined by protein-sequence coverage and spectral counting in MudPIT experiments) — resulted in greatly reduced amounts of endoplasmic-reticulum-associated and cell-surface-associated ΔF508-CFTR (Fig. 3c). Overexpression of these co-chaperones, however, had opposite effects on ΔF508-CFTR, with an increase in p23 leading to more endoplasmic-reticulum-associated CFTR and an increase in FKBP8 leading to less. These data were thought to reflect the distinct roles of p23 and FKBP8 in modulating specific aspects of the HSP90-guided folding cycle. Notably, overexpression of either co-chaperone failed to increase the amount of cell-surface-associated ΔF508-CFTR (or wild-type CFTR), suggesting that they modulate the initial folding and stability of CFTR but do not participate in the subsequent steps that are required to deliver the folded channel to the endoplasmic-reticulum export machinery.

RNAi-mediated knockdown of a third HSP90 co-chaperone present in wild-type and ΔF508-CFTR immunoprecipitates, AHA1, substantially corrected the amount of both endoplasmic-reticulum-associated and cell-surface-associated ΔF508-CFTR (Fig. 3c). These data suggest that disruption of AHA1, unlike p23 or FKBP8, facilitates a folding pathway that favours not only stability of the channel but also coupling to the endoplasmic-reticulum export machinery. Potentially consistent with this premise, a decrease in CFTR-bound HSP90 was observed in cells in which AHA1 had been knocked down, similar to the finding from MudPIT analyses that wild-type CFTR immunoprecipitates contained less HSP90 than ΔF508-CFTR immunoprecipitates. Collectively, these data

indicate that a reduction in the amount of AHA1 might alter the kinetics of HSP90–CFTR interactions, thereby increasing the efficiency of transition from folding to export pathways. Finally, the authors showed that a considerable proportion of the cell-surface-associated ΔF508-CFTR channels were functional, as determined by chloride conductance measurements.

Mass-spectrometry-based proteomic studies of the CFTR interactome thus identified specific co-chaperone and chaperone folding pathways that seem to control mutant channel stability, cell-surface expression and function. Why might the basal chaperone machinery of a cell, or 'chaperome', prevent proper assembly and transport of ΔF508-CFTR, thereby exacerbating the disease phenotype? The authors[37] speculate that the folding energetics of ΔF508-CFTR lie outside the capacity of the normal chaperome environment, which has been evolutionarily optimized to fold wild-type proteins (and to eliminate misfolded proteins). A provocative extension of this idea is that genetic or pharmacological interventions that shift the chaperome so that it can support the folding and transport of mutant proteins could be used to treat patients with cystic fibrosis, as well as those with other protein-conformational disorders.

Functional characterization of protein pathways

One of the original and most enduring applications of mass-spectrometry-based proteomics is the comparative analysis of biological samples that differ in specific physiological or pathophysiological phenotypes (that is, comparing 'disease' and 'normal'[39]). These studies are intended to identify the minimal protein 'signatures' that depict and, ideally, determine the higher-order biological processes under investigation. As highlighted in this section, mass-spectrometry-based proteomics carried out in this comparative analysis mode has successfully identified previously unknown protein pathways with key roles in a wide range of biological systems.

Kinase pathways that regulate sex-specific functions in *Plasmodium*

Malaria is caused by unicellular parasites of the genus *Plasmodium*. These parasites undergo a complex series of highly regulated life-cycle transitions that allow transmission between vertebrates and mosquitoes[40]. Chief among these life-cycle stages is the generation of haploid sexually differentiated (male and female) cells, termed gametocytes. In vertebrate blood, gametocytes are in a state of arrest, but on transfer to the mosquito mid-gut, they become activated and differentiate into gametes, which fertilize and, eventually, produce infectious oocysts. Despite the importance of sexual development to the transmission of *Plasmodium* spp., the proteins that distinguish male and female cells have not been systematically inventoried. Andrew Waters and colleagues set out to address this important problem through an innovative combination of advanced cell-biological models and proteomic technologies[41].

Previous efforts to characterize sex-specific proteins had been confounded by a technical inability to separate and purify male and female gametocytes. Waters and colleagues overcame this difficulty by creating transgenic lines of *Plasmodium berghei* that produce green fluorescent protein (GFP) under the control of a male-specific or a female-specific promoter (from the genes encoding α-tubulin II and elongation factor 1α, respectively) (Fig. 4a). These lines enabled male gametocytes and female gametocytes to be selectively enriched by flow cytometry (Fig. 4b). These sex-specific cell populations were then compared with one another (and with gametocyte-free (asexual) blood stages) by mass-spectrometry-based proteomics. Specifically, proteomes were separated into ten fractions by one-dimensional SDS–PAGE, and each fraction was digested with trypsin and analysed by LC–MS/MS. Sex-specific proteins were identified by comparing the number of unique 'tryptic peptides' in each sample.

A remarkable number of sex-specific proteins were identified: there were 236 unique proteins in male gametocytes, and 101 in female gametocytes (Fig. 4c). Analysis of the sex-specific proteomes showed clear enrichment for protein families that are functionally linked to male-gametocyte and female-gametocyte biology. For example, nearly 70% of the *Plasmodium* proteins that are annotated as DNA-replication proteins (17 of 25) were highly represented in male gametocytes, which is consistent with the more extensive genome replication that these cells undergo during the gamete-activation cycle. The male-gametocyte proteome was also strongly enriched in axoneme proteins, which form the flagella required for motility of male gametes. The female-gametocyte proteome, by contrast, contained larger amounts of ribosomal and mitochondrial proteins than the male-gametocyte proteome.

Waters and colleagues next selected individual sex-specific proteins for functional analysis, to determine whether they have important

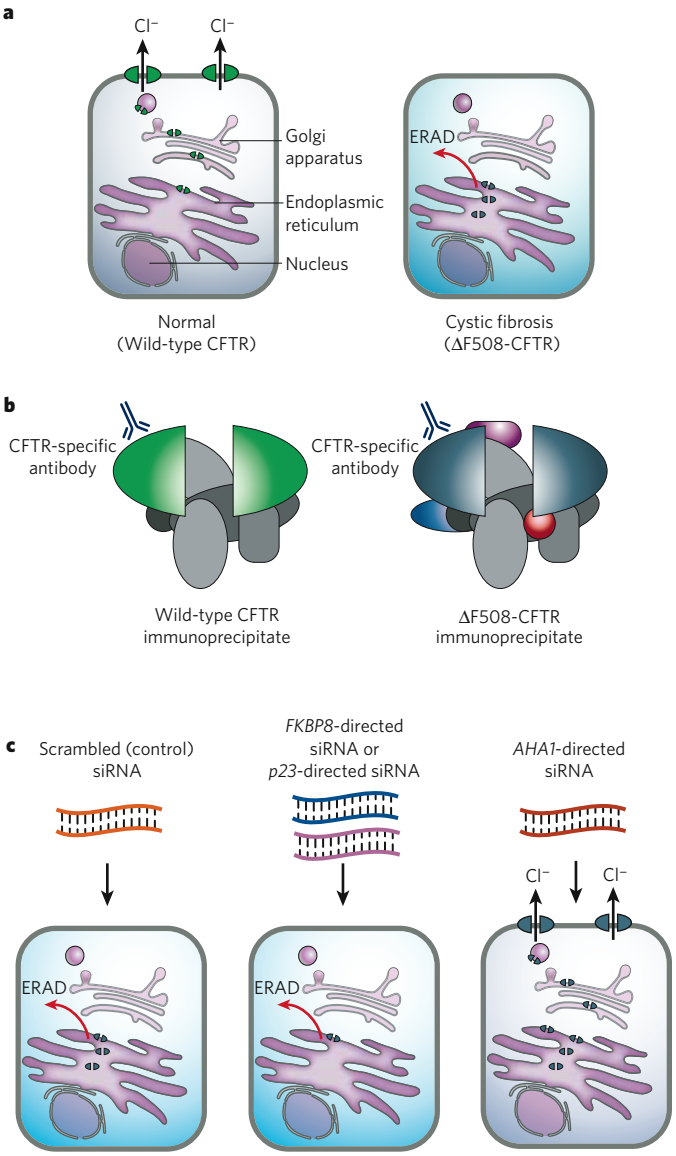


Figure 3 | Discovery of chaperone complexes that regulate CFTR folding and endoplasmic-reticulum-mediated export. a, The cellular fate of CFTR chloride channels is depicted. Wild-type CFTR is transported to the plasma membrane. By contrast, ΔF508-CFTR, the mutant protein present in individuals with cystic fibrosis, is degraded by endoplasmic-reticulum-mediated pathways (ERAD) before it reaches the plasma membrane. **b**, Immuno-enrichment of proteins bound to wild-type CFTR and ΔF508-CFTR identified several chaperones and co-chaperones[37]. **c**, RNAi-mediated knockdown of these chaperones and co-chaperones was carried out, using small interfering RNA (siRNA) directed against the corresponding mRNAs. In the case of p23 and FKBP8 (centre), less ΔF508-CFTR was found in the endoplasmic reticulum, indicating that these proteins regulate the folding and stability of CFTR proteins in the endoplasmic reticulum. By contrast, in the case of AHA1 (right), more ΔF508-CFTR was found both in the endoplasmic reticulum and at the cell surface, indicating that this chaperone controls both the folding of CFTR proteins and their export from the endoplasmic reticulum[37].

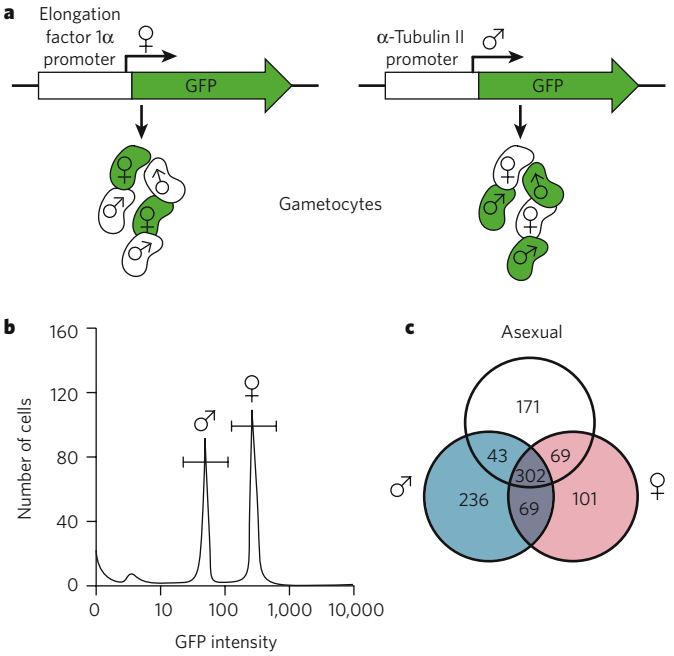


Figure 4 | Identification of male-gametocyte-specific and female-gametocyte-specific proteins in *Plasmodium berghei*. **a**, Transgenic *Plasmodium berghei* lines that produce GFP under the control of a female-gametocyte-specific promoter (from the gene encoding elongation factor 1α) or a male-gametocyte-specific gene promoter (from the gene encoding α-tubulin II) were generated[41]. **b**, Enriched populations of female and male gametocytes were then obtained by flow cytometry, gating on GFP signals (shown for the line in which female gametocytes produce GFP). **c**, Comparative proteomic analysis of male-gametocyte-enriched, female-gametocyte-enriched and asexual populations identified 236, 101 and 171 proteins, respectively, that are expressed solely in each cell type. Among these proteins were identified two protein kinases that contribute to male-gametocyte-specific and female-gametocyte-specific cellular functions.

roles in male-gametocyte or female-gametocyte biology. The authors focused on two protein kinases: mitogen-activated protein kinase 2 (MAP2; accession number PB000659.00.0), which was found only in male gametocytes; and NIMA-related kinase (NEK4; accession number PB001094.00.0), which was found only in female gametocytes. Targeted disruption of the gene encoding MAP2 resulted in male gametocytes that can re-enter the cell cycle after activation and complete genome replication but fail to enter nuclear division. Disruption of the gene encoding NEK4, by contrast, did not seem to impair gamete formation but, instead, arrested zygote development. Cross-fertilization studies confirmed that the latter phenotype was due to defective female (but not male) gametes.

In summary, developing an innovative cell-biological strategy to enrich distinct sexual stages of the *P. berghei* life cycle allowed the generation of high-quality cellular models for in-depth analysis by mass-spectrometry-based proteomics. The output of the proteomic investigations was the most comprehensive inventory of sex-specific parasite proteins generated so far, including the discovery of novel protein kinases that regulate male-specific and female-specific signalling pathways. Interestingly, both MAP2 and NEK4 belong to protein-kinase subfamilies (the MAP and NEK subfamilies) that have multiple members in *Plasmodium* spp.[42]. These studies are therefore a compelling example of the value of comparative proteomics for assigning unique (that is, non-redundant) cellular functions to uncharacterized members of protein classes.

An ether-lipid signalling pathway that supports tumour pathogenesis

Cancer cells have long been suspected to have alterations in metabolism that support their malignant behaviour. Most cancer cells, for example, have a greater dependence on glycolysis than on oxidative phosphorylation for energy production, a phenomenon referred to as the Warburg effect[22]. In an effort to map dysregulated biochemical pathways in cancer more globally, Benjamin Cravatt and colleagues used a chemical proteomic technology known as activity-based protein profiling (ABPP)[43], in conjunction with mass-spectrometry-based analytical platforms (such as MudPIT), to identify enzyme activities that are increased in aggressive cancer cell lines and primary tumours in humans[44].

In ABPP, active-site-directed probes are used to profile the functional state of enzymes directly in native proteomes[43]. ABPP probes contain two main elements: a reactive group that binds to, and covalently labels, many enzymes from a given mechanistic class; and a reporter group, such as biotin or a fluorophore, that allows detection, enrichment and identification of probe-modified enzymes (Fig. 5a). In their initial studies, Cravatt and colleagues used fluorophosphonate-containing ABPP probes[45,46] to profile the activities of serine hydrolases in a panel of human cancer cell lines[44]. These experiments identified sets of enzyme activities that distinguished cancer cells on the basis of tissue of origin and state of aggressiveness. Chief among these enzymes was a previously uncharacterized transmembrane enzyme KIAA1363 (also known as AADACL1), increased amounts of which were found in aggressive lines from several tumour types, including breast cancer, ovarian cancer and melanoma (Fig. 5b). Cravatt and colleagues later showed by ABPP–MudPIT analysis that the activity of KIAA1363 is much higher in oestrogen-receptor-negative primary breast tumours from humans than in oestrogen-receptor-positive primary breast tumours, which are usually less aggressive, or in normal breast tissue[47].

Cravatt and colleagues next used a competitive version of ABPP[48] to develop a potent and selective inhibitor of KIAA1363, which they named AS115 (Fig. 5c). Treatment of cancer cells with this inhibitor, followed by metabolomic analysis using untargeted LC–MS methods[49], revealed that KIAA1363 regulates an unusual class of neutral lipids: the monoalkylglycerol ethers (MAGEs)[50] (Fig. 5d). Additional studies confirmed that KIAA1363 is a 2-acetyl-MAGE hydrolase, producing large amounts of MAGEs in aggressive cancer cells. These MAGEs are, in turn, converted into biologically active lysophospholipids, such as lysophosphatidic acid (LPA). By contrast, inhibiting KIAA1363 stabilizes 2-acetyl-MAGE, resulting in its conversion into another class of signalling molecule, the lipid platelet-activating factor. Finally, RNAi-mediated knockdown of the protein and activity of KIAA1363 led to a marked decrease in the amount of MAGE and LPA lipids in cancer cells, correlating with significant reductions in the migratory and tumour-forming potential of these cells (Fig. 5d).

In summary, Cravatt and colleagues used a combination of mass-spectrometry-based functional proteomic and metabolomic methods to determine that the enzyme KIAA1363 is more abundant and has a higher activity in aggressive cancer cells, where it is a key node that bridges platelet-activating factor and LPA in an ether-lipid signalling network. Considering that disruption of this network impaired cancer-cell migration and tumour growth, the KIAA1363–ether-lipid pathway probably has a key role in regulating important aspects of cancer pathogenesis.

Phosphoprotein networks involved in the DNA-damage response

Post-translational modifications constitute one of the most pervasive mechanisms for regulating protein function in cells and tissues. Protein phosphorylation, in particular, dynamically modulates numerous signalling pathways and is controlled by the complementary action of protein kinases and protein phosphatases. One of the big challenges in the post-genomic era is determining the endogenous substrates of the more than 500 protein kinases in the human proteome[51]. Recently, Stephen Elledge and colleagues[52] introduced a creative mass-spectrometry-based proteomic strategy that allowed them to make a comprehensive inventory of substrates for the protein kinases ATM (ataxia telangiectasia mutated) and ATR (ATM and Rad3 related), which are involved in the DNA-damage-response pathway[52].

Previous studies had identified about 25 ATM and/or ATR substrates, which contained an unusual consensus sequence for phosphorylation: Ser/Thr-Gln[53]. On the basis of this information, Elledge and colleagues used a panel of 68 antibodies specific for phospho-Ser-Gln or

phospho-Thr-Gln to immunoprecipitate candidate substrates for ATM and ATR from human cells that were either exposed to ionizing radiation to induce the DNA-damage response or not irradiated (Fig. 6). These cell populations had previously been subjected to stable isotope labelling[54,55], so radiation-induced phosphorylation events could be quantified by LC–MS/MS analysis. Relative quantification of heavy-isotope-labelled and light-isotope-labelled phosphopeptide pairs identified 905 phosphorylation sites, across 700 proteins, that had fourfold higher signals in irradiated cells than in non-irradiated cells. Thus, in a single set of experiments, the researchers increased the number of candidate ATM and ATR substrates by more than 20-fold (from about 25 proteins to 700 proteins). The increase in phosphorylation found in irradiated cells was confirmed for several candidate substrates by immunoblotting with antibodies specific for phospho-Ser-Gln or phospho-Thr-Gln.

The researchers next examined whether the newly identified substrates have a role in the DNA-damage response, by systematically disrupting expression of the corresponding genes by RNAi. Of the 37 substrates examined, 35 were found to contribute to at least one aspect of the DNA-damage response. Although these studies do not directly test whether the phosphorylation state of the proteins is crucial for their function, the results indicate that many more proteins contribute to the DNA-damage response than was originally thought, and these proteins are subject to dynamic phosphorylation in response to DNA-damage signals. Interestingly, there were several cases in which multiple components of a given pathway were phosphorylated, leading the authors to conclude that protein kinases can increase their effect on specific signalling pathways by simultaneously phosphorylating several nodes.

The authors then rapidly mined their phosphoproteomic data sets, facilitating the functional annotation of two previously uncharacterized proteins. One of these proteins, which the authors named abraxas (also known as CCDC98 and FLJ13614), was identified as a potential ATM and/or ATR substrate. Abraxas was more heavily phosphorylated in irradiated cells than in non-irradiated cells, and it formed a complex with RAP80 (also known as UIMC1) and BRCA1, which was required for resistance to DNA damage, control of the cycle checkpoint at the G2–M boundary and repair of DNA[56]. The other protein, which the authors named FANCI (also known as KIAA1794), was found to form a complex with FANCD2, which then localized to chromatin in response to DNA damage[57]. Interestingly, a mutation in the gene encoding FANCI had been causally linked to Fanconi's anaemia, a syndrome that impairs development and increases the risk of developing cancer.

These studies, together with others[58–61], underscore the rapid development of quantitative mass-spectrometry methods for mapping protein phosphorylation sites in proteomes. Similar approaches are emerging for the global analysis of other key protein modifications, including acetylation[62,63], methylation[63,64], glycosylation[65] and ubiquitylation[66]. We expect that these methods will also help to improve our understanding of the role of post-translational modifications in regulating protein function in biological systems.

Insulin pathways in *Caenorhabditis elegans* dauer formation and ageing

Genetic studies in *C. elegans* have determined that the signalling pathway involving insulin and insulin-like growth factor has an important role in regulating lifespan[67]. For example, disruption of the *C. elegans* receptor DAF-2, which is homologous to the mammalian receptors for both insulin and insulin-like growth factor 1, extends lifespan and increases entry to the dauer phase (a phase characterized by delayed development, which *C. elegans* can enter if environmental conditions are unfavourable early in development)[68]. To understand the molecular basis of these marked changes in physiology, John Yates and colleagues carried out a quantitative mass-spectrometry-based proteomic analysis of wild-type and *daf-2* mutant strains of *C. elegans*[69].

Two forms of quantification were used: ratiometric analysis of proteomes from both wild-type *C. elegans* and *daf-2* mutants with a reference proteome corresponding to wild-type *C. elegans* fed on ^{15}N-enriched bacteria[70]; and direct spectral counting[71] of unlabelled proteins in wild-type and *daf-2* mutant proteomes (Fig. 7a). Together, these methods identified 86 proteins that were differentially expressed in *daf-2* mutants, 47 that were more abundant and 39 that were less abundant than in wild-type *C. elegans*. There were good correlations between the proteomic data obtained with the two methods, indicating that either approach can provide an accurate estimate of the relative levels of proteins in two or more biological samples. The authors verified their proteomic data by selecting several proteins from wild-type strains and *daf-2* mutant strains for analysis by immunoblotting.

Interestingly, proteins that had similar changes in abundance in the *daf-2* mutant strain tended to show a functional relationship. For example, as a group, the more abundant proteins tended to have translation-elongation and lipid-transport functions, whereas the less-abundant proteins were over-represented in the categories of amino-acid biosynthesis, reactive-oxygen-species metabolism and carbohydrate metabolism. Yates and colleagues next tested whether these proteins

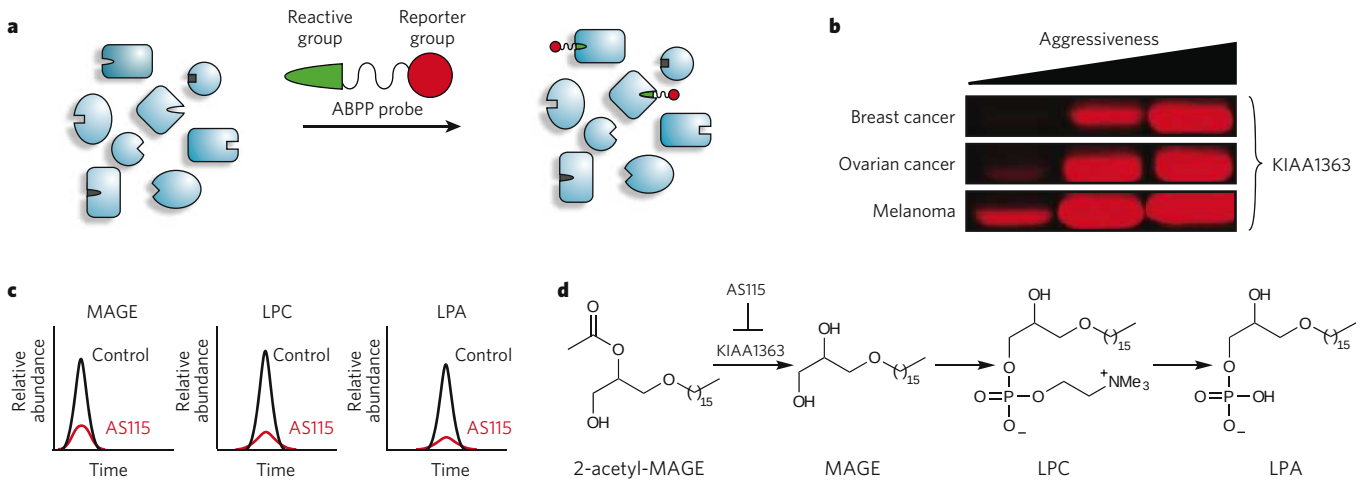


Figure 5 | Discovery of an ether-lipid signalling pathway that supports cancer pathogenesis. a, The general structure and mechanism of action of ABPP probes are shown, with proteins of various activities in blue and a probe with a specific reactive group. **b**, ABPP of a panel of human tumour cell lines identified an uncharacterized hydrolase, KIAA1363, in aggressive cell lines from several tumour types. The activity of KIAA1363 increased with the aggressiveness of the cell lines (as determined by in-gel fluorescence scanning of probe-labelled KIAA1363)[44]. **c**, Inactivation of KIAA1363 by the selective inhibitor AS115 (or short hairpin RNA probes) decreased the abundance of a family of ether lipids, including MAGEs and alkyl-lysophospholipids (lysophosphatidylcholine (LPC) and LPA), as determined by LC–MS analysis[50]. **d**, These results suggest a model in which KIAA1363 regulates an ether-lipid pathway that proceeds from MAGEs to LPC and LPA. Disruption of this lipid network by blockade of KIAA1363 inhibited cancer-cell migration and tumour growth (not shown)[50]. Me, methyl.

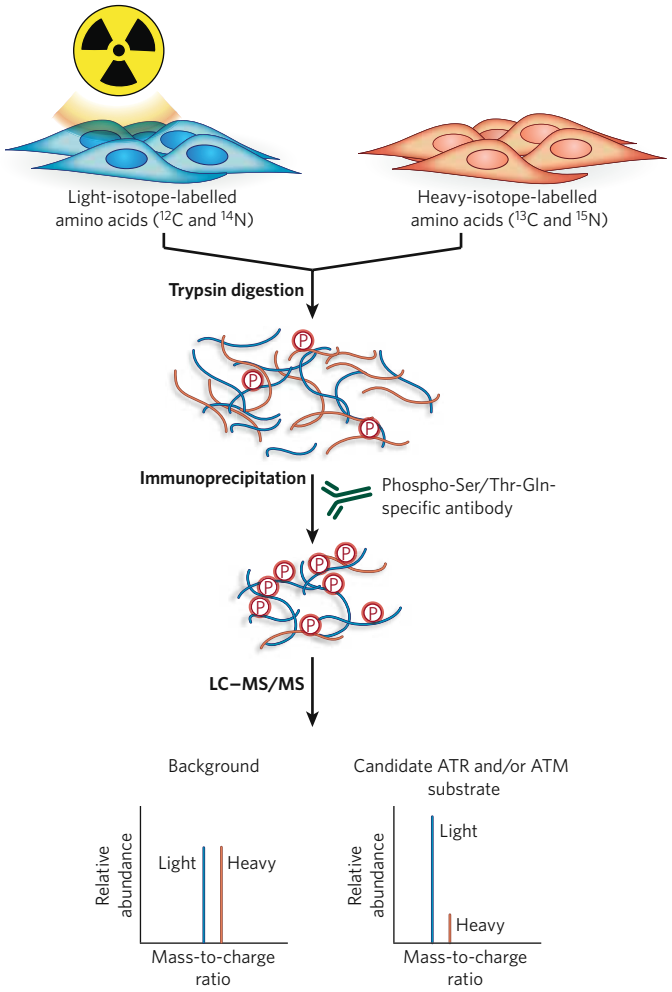


Figure 6 | Identification of candidate ATM and/or ATR substrates involved in the DNA-damage response. Cells treated with light-isotope-labelled amino acids were exposed to ionizing radiation, and cells treated with heavy-isotope-labelled amino acids were maintained under control conditions. Candidate ATM and ATR substrates were then identified by trypsin digestion of whole-cell proteomes, followed by immunoprecipitation with antibodies specific for the consensus ATM and ATR phosphorylation motif phospho-Ser/Thr-Gln, and then LC–MS/MS analysis. Phosphoproteins produced in response to irradiation were identified by ratiometric analysis of mass signals from light-isotope-labelled cells and heavy-isotope-labelled cells. Many of these proteins were found to have important roles in the DNA-damage response[52].

affected DAF-2-dependent processes such as lifespan. Curiously, RNAi-mediated knockdown in wild-type *C. elegans* of the mRNAs encoding proteins that had increased abundance in *daf-2* mutants tended to extend the lifespan further, whereas knockdown of the mRNAs encoding less-abundant proteins shortened the lifespan of wild-type *C. elegans* (Fig. 7b). These results suggest that many of the proteomic changes observed in *daf-2* mutants reflect compensatory changes in metabolic and/or signalling pathways that limit the impact of loss of DAF-2 function.

Principal among the observed compensatory pathways was TAX-6 (also known as CNA-1), the *C. elegans* orthologue of the protein phosphatase known as calcineurin A. Significantly more TAX-6 was present in *daf-2* mutants than in wild-type *C. elegans*. In addition, disruption of the gene encoding TAX-6 (*tax-6*) produced a similar phenotype to loss of DAF-2 (that is, extended lifespan and increased entry to the dauer phase). Disruption of both *tax-6* and *daf-2* resulted in even more marked phenotypes. Collectively, these data indicate that TAX-6 is part of a feedback loop that buffers the effects of DAF-2 on longevity, through compensatory mechanisms (Fig. 7c). A provocative extension of this idea is that pharmacological strategies to block such compensatory pathways might be useful for extending the lifespan of animals. From a more technical perspective, this study, together with another study by Yates and colleagues[72], shows that stable isotope labelling can be applied to intact organisms, as well as to cell-culture preparations, thus greatly expanding the potential applications of this quantitative mass-spectrometry-based proteomic method.

Emergent themes for mass-spectrometry-based proteomics

The studies described in this review have several common conceptual and experimental themes that are instructive for researchers interested in using mass-spectrometry-based proteomics. First, it is clear that, to ask specific biological questions, well-configured model systems need to be established. Proteomic experiments produce large amounts of data. For these data sets to deliver answers or inspire compelling hypotheses that explain the molecular basis of complex biological processes, well-designed experimental systems and controls must be incorporated into the research plan. Not surprisingly, experimental systems often involve pathophysiological states for which clinical phenotypes are well described. Using the appropriate controls allows investigators rapidly to winnow down proteomic observations to a manageable number of proteins that show changes in abundance, activity or post-translational modification in the experimental model under study.

If carried out properly, mass-spectrometry-based proteomics experiments should uncover a set of proteins associated with a specific cellular or physiological process. Testing the function of these proteins, however, requires 'targeted' follow-up studies that use complementary experimental approaches. A second theme is the emergence of RNAi as a near-universal method to perturb the production of any protein in cells and organisms, offering researchers a powerful strategy to test the function of proteins identified in proteomic experiments. RNAi also has the advantage of operating on a scale that is compatible with screening the biological function of hundreds to thousands of candidate proteins[73,74], making it an attractive method to rapidly validate targets discovered in large-scale proteomic endeavours. Perhaps the best way to picture the growing synergistic relationship between mass-spectrometry-based proteomic techniques and RNAi techniques is to view the former approach as a hypothesis-generating engine and the latter as a tool for testing these hypotheses. In this manner, proteomic observations can be connected to function or phenotype.

A third common theme among the studies highlighted here is that the follow-up biological experiments were carried out by the same research group as the original mass-spectrometry-based proteomics investigation. Although repositories of proteomic data are undoubtedly useful, this finding suggests that the primary biological users of proteomic information are typically the generators of these data. There are several reasons why this might be the case. First, biologists are inundated with large-scale data sets, including those that inventory transcript, protein and metabolite expression, as well as protein–protein interactions and post-translational modification state. This glut of molecular information almost certainly has a saturating effect on potential users, who may face too many candidate targets or pathways to explore. Second, potential users might be concerned about the quality of mass-spectrometry-based proteomic data (for example, the number of false-positive and false-negative results). Follow-up biological studies are not trivial in terms of cost or time, and having confidence in the quality of the data would probably lower the 'activation energy barrier' for secondary users of proteomic results. Last, it might simply take more time for secondary users to incorporate mass-spectrometry-based proteomic data sets into their biological studies, or secondary users might incorporate data from proteomic experiments mainly to validate observations from their own experiments. Thus, there might be particular issues to overcome before repositories of large-scale proteomic data influence hypothesis-driven research, which often involves highly specific objectives for which proteomic data might be too general to address. This situation should improve as new methods for mining stored proteomic data are developed. It should also be noted that it is much easier to track scientific

progress if a common authorship is preserved. We might therefore be underestimating the number of researchers who have capitalized on repositories of mass-spectrometry-based proteomic data to gain new insights into biological systems.

We have highlighted experimental commonalities among mass-spectrometry-based proteomic studies that made important biological discoveries; however, there are also some noteworthy differences. For example, several methods for protein quantification have been used: these include ICAT[25]; stable isotope labelling of cells[52] and organisms[69]; and label-free techniques such as unique peptide number[41], protein-sequence coverage[37] and spectral counting[37,47,69]. Given that each of these strategies is generally successful, does there need to be a single form of data collection and analysis in quantitative mass-spectrometry-based proteomic experiments? This question can be distilled to the issue of balancing accuracy and ease of implementation. Label-free methods are the simplest and most cost-effective to carry out, but they lack the precision of isotope-labelling techniques. However, as long as researchers are committed to validating a portion of their proteomic results by using complementary techniques (for example, immunoblotting or selective-reaction monitoring), confidence in the overall data sets acquired with either method should be achievable. These validation experiments should, for example, readily identify false-positive data, which can be eliminated from further analysis. False-negative results (that is, changes that occur but are not detected) are more problematic but are almost certainly minimized as the accuracy of the quantification method increases. On this note, the discovery of Tfb5 as a component of TFIIH is worth revisiting. This protein showed only a twofold increase in ICAT signal in enriched TFIIH complexes[25], a signal difference that probably would not have registered as meaningful if less-accurate (label-free) quantification methods had been used. Regardless, in all of the studies highlighted here, the overall importance of the proteomic data sets was established by follow-up biological experiments.

Conclusions and future directions

Future technical challenges for mass-spectrometry-based proteomics mainly relate to the nature of proteins in biological systems. Proteins have a wide range of abundances, and this is further confounded by the myriad post-translational modifications that are dynamically regulated by cellular context and time. To capture the various states of proteins in a cell fully, proteomes must therefore be sampled in different conditions and at several time points following perturbation. Several technical aspects of mass spectrometers need to be improved to meet the demands for higher throughput and proteome coverage without sacrificing information content. First, advances in instrument scan speed would allow more frequent sampling of ions. Higher rates of sampling would translate into more tandem mass spectra acquired per unit time, which would, in turn, enable higher-resolution chromatography methods to be used. Increased sampling rates should also improve dynamic range, because lower-abundance ions are more likely to be detected. Second, coupling these changes to continued improvements in sensitivity and mass accuracy measurements, the gain in dynamic range could be multiplied. Increased resolution and mass accuracy should also strengthen confidence in peptide identifications and facilitate the discovery of protein modifications. Third, advances in 'top-down' mass spectrometry for sequence-based characterization of intact proteins can allow patterns of modifications on a protein to be correlated with specific activities or functions. At present, top-down mass spectrometry is most effective for small proteins (< 25 kDa) and presents difficulties for analysing larger proteins[75]. Key areas for the improvement of top-down mass spectrometry are the development of more general fragmentation methods for large proteins, and of higher-throughput and more-robust methods to introduce intact proteins into the mass spectrometer. Final issues to consider relate to the throughput and sample demands of standard mass-spectrometry-based proteomics experiments. Unbiased, global methods such as a two-dimensional liquid-chromatography-based shotgun proteomics require considerable time (several hours per sample) and material (> 0.1 mg protein per sample). Using other strategies such as accurate mass tagging and single-ion reaction monitoring of peptides can increase throughput and reduce sample demands, but they limit the analysis to peptides or proteins that are known to be present in a mixture and, therefore, preclude serendipitous discoveries[76,77]. Because mass-spectrometry-based proteomic methodology continues to develop at a rapid pace, there is much hope that these and other problems will be solved.

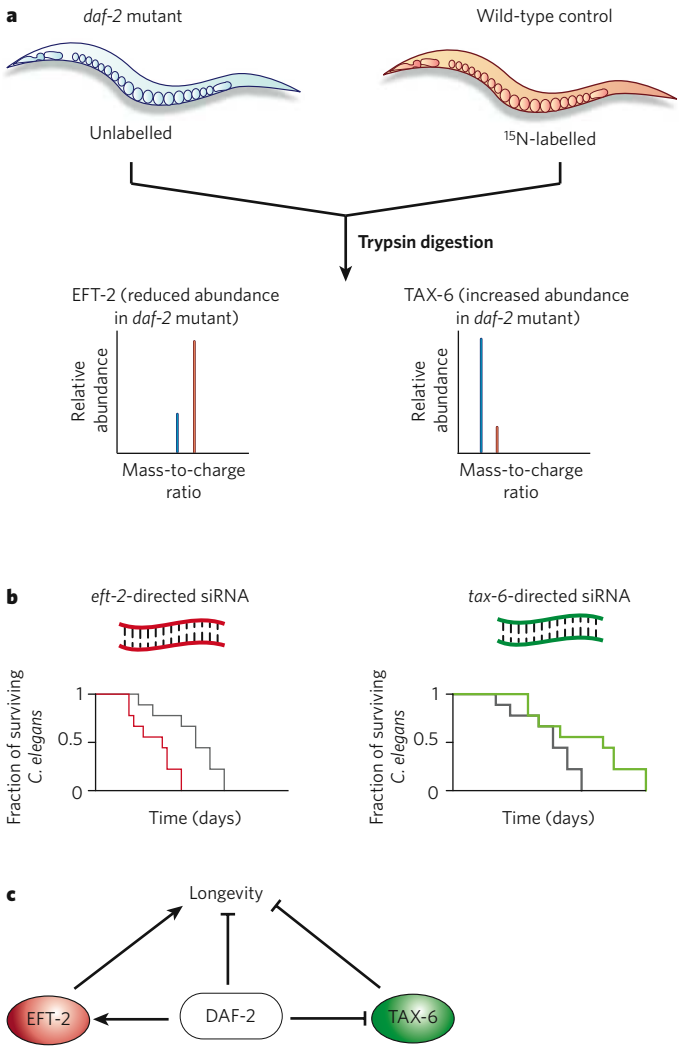


Figure 7 | Discovery of DAF-2-regulated protein pathways that modulate longevity in *Caenorhabditis elegans*. **a**, A quantitative proteomic analysis of changes in protein abundance was carried out in *daf-2* mutant *C. elegans*, by using metabolic labelling and MudPIT analysis[69]. Shown are representative examples of proteins that either decreased (EFT-2) or increased (TAX-6) in abundance in *daf-2* mutants. **b**, Follow-up studies on differentially expressed proteins identified cases in which RNAi-mediated knockdown of the corresponding mRNAs decreased (EFT-2) or increased (TAX-6) the lifespan of worms. These results suggest that the proteins participate in compensatory pathways that limit the effects of *daf-2* mutation on longevity and dauer formation. **c**, A model of how DAF-2-regulated proteins participate in compensatory pathways that affect longevity was assembled from the results of these experiments.

It is clear that biologists are becoming increasingly savvy users of mass-spectrometry instrumentation and, conversely, that mass spectrometrists are gaining familiarity with other biological techniques. We therefore expect that distinctions between these types of scientist will soon begin to lose meaning. How long might it be, for example, before mass spectrometers stand alongside centrifuges and PCR machines as core pieces of equipment in every biology lab? Wouldn't that be the ultimate sign of biological impact for this powerful analytical technology? ■

1. Whisstock, J. C. & Lesk, A. M. Prediction of protein function from protein sequence and structure. *Q. Rev. Biophys.* **36,** 307–340 (2003).
2. Galperin, M. Y. & Koonin, E. V. 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic Acids Res.* **32,** 5452–5463 (2004).
3. Zhu, H., Bilgin, M. & Snyder, M. Proteomics. *Annu. Rev. Biochem.* **72,** 783–812 (2003).
4. de Hoog, C. L. & Mann, M. Proteomics. *Annu. Rev. Genomics Hum. Genet.* **5,** 267–293 (2004).
5. MacBeath, G. Protein microarrays and proteomics. *Nature Genet.* **32** (suppl.), 526–532 (2002).
6. Hall, D. A., Ptacek, J. & Snyder, M. Protein microarray technology. *Mech. Ageing Dev.* **128,** 161–167 (2007).
7. Causier, B. Studying the interactome with the yeast two-hybrid system and mass spectrometry. *Mass Spectrom. Rev.* **23,** 350–367 (2004).
8. Stevens, R. C., Yokoyama, S. & Wilson, I. A. Global efforts in structural genomics. *Science* **294,** 89–92 (2001).
9. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422,** 198–207 (2003).
10. Yates, J. R. Mass spectral analysis in proteomics. *Annu. Rev. Biophys. Biomol. Struct.* **33,** 297–316 (2004).
11. Domon, B. & Aebersold, R. Mass spectrometry and protein analysis. *Science* **312,** 212–217 (2006).
12. Andersen, J. S. & Mann, M. Organellar proteomics: turning inventories into insights. *EMBO Rep.* **7,** 874–879 (2006).
13. Cusick, M. E., Klitgord, N., Vidal, M. & Hill, D. E. Interactome: gateway into systems biology. *Hum. Mol. Genet.* **14,** R171–R181 (2005).
14. Michnick, S. W. Proteomics in living cells. *Drug Discov. Today* **9,** 262–267 (2004).
15. Neubauer, G. *et al.* Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. *Nature Genet.* **20,** 46–50 (1998).
16. Wang, Y. *et al.* BASC, a super complex of BRCA1-associated proteins involved in the recognition and repair of aberrant DNA structures. *Genes Dev.* **14,** 927–939 (2000).
17. Rout, M. P. *et al.* The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J. Cell Biol.* **148,** 635–651 (2000).
18. Bouwmeester, T. *et al.* A physical and functional map of the human TNF-α/NF-κB signal transduction pathway. *Nature Cell Biol.* **6,** 97–105 (2004).
19. Das, R. *et al.* SR proteins function in coupling RNAP II transcription to pre-mRNA splicing. *Mol. Cell* **26,** 867–881 (2007).
20. Danial, N. N. *et al.* BAD and glucokinase reside in a mitochondrial complex that integrates glycolysis and apoptosis. *Nature* **424,** 952–956 (2003).
21. Harada, H. *et al.* Phosphorylation and inactivation of BAD by mitochondria-anchored protein kinase A. *Mol. Cell* **3,** 413–422 (1999).
22. Gatenby, R. A. & Gillies, R. J. Why do cancers have high aerobic glycolysis? *Nature Rev. Cancer* **4,** 891–899 (2004).
23. Wang, X. The expanding role of mitochondria in apoptosis. *Genes Dev.* **15,** 2922–2933 (2001).
24. Coulombe, B., Jeronimo, C., Langelier, M. F., Cojocaru, M. & Bergeron, D. Interaction networks of the molecular machines that decode, replicate, and maintain the integrity of the human genome. *Mol. Cell. Proteomics* **3,** 851–856 (2004).
25. Ranish, J. A. *et al.* Identification of TFB5, a new component of general transcription and DNA repair factor IIH. *Nature Genet.* **36,** 707–713 (2004).
26. Gygi, S. P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnol.* **17,** 994–999 (1999).
27. Cenkci, B., Petersen, J. L. & Small, G. D. REX1, a novel gene required for DNA repair. *J. Biol. Chem.* **278,** 22574–22577 (2003).
28. Giglia-Mari, G. *et al.* A new, tenth subunit of TFIIH is responsible for the DNA repair syndrome trichothiodystrophy group A. *Nature Genet.* **36,** 714–719 (2004).
29. Vermeulen, W. *et al.* Sublimiting concentration of TFIIH transcription/DNA repair factor causes TTD-A trichothiodystrophy disorder. *Nature Genet.* **26,** 307–313 (2000).
30. Krogan, N. J. *et al.* Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440,** 637–643 (2006).
31. Krebs, M. P., Noorwez, S. M., Malhotra, R. & Kaushal, S. Quality control of integral membrane proteins. *Trends Biochem. Sci.* **29,** 648–655 (2004).
32. Kelly, J. W. & Balch, W. E. The integration of cell and chemical biology in protein folding. *Nature Chem. Biol.* **2,** 224–227 (2006).
33. Riordan, J. R. Assembly of functional CFTR chloride channels. *Annu. Rev. Physiol.* **67,** 701–718 (2005).
34. Qu, B. H., Strickland, E. H. & Thomas, P. J. Localization and suppression of a kinetic defect in cystic fibrosis transmembrane conductance regulator folding. *J. Biol. Chem.* **272,** 15739–15744 (1997).
35. Loo, M. A. *et al.* Perturbation of Hsp90 interaction with nascent CFTR prevents its maturation and accelerates its degradation by the proteasome. *EMBO J.* **17,** 6879–6887 (1998).
36. Meacham, G. C., Patterson, C., Zhang, W., Younger, J. M. & Cyr, D. M. The Hsc70 co-chaperone CHIP targets immature CFTR for proteasomal degradation. *Nature Cell Biol.* **3,** 100–105 (2001).
37. Wang, X. *et al.* Hsp90 cochaperone Aha1 downregulation rescues misfolding of CFTR in cystic fibrosis. *Cell* **127,** 803–815 (2006).
38. Washburn, M. P., Wolters, D. & Yates, J. R. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnol.* **19,** 242–247 (2001).
39. Hanash, S. Disease proteomics. *Nature* **422,** 226–232 (2003).
40. Kumar, N. *et al.* Molecular complexity of sexual development and gene regulation in *Plasmodium falciparum*. *Int. J. Parasitol.* **34,** 1451–1458 (2004).
41. Khan, S. M. *et al.* Proteome analysis of separated male and female gametocytes reveals novel sex-specific *Plasmodium* biology. *Cell* **121,** 675–687 (2005).
42. Ward, P., Equinet, L., Packer, J. & Doerig, C. Protein kinases of the human malaria parasite *Plasmodium falciparum*: the kinome of a divergent eukaryote. *BMC Genomics* **5,** 79 (2004).
43. Jessani, N. & Cravatt, B. F. The development and application of methods for activity-based protein profiling. *Curr. Opin. Chem. Biol.* **8,** 54–59 (2004).
44. Jessani, N., Liu, Y., Humphrey, M. & Cravatt, B. F. Enzyme activity profiles of the secreted and membrane proteome that depict cancer invasiveness. *Proc. Natl Acad. Sci. USA* **99,** 10335–10340 (2002).
45. Liu, Y., Patricelli, M. P. & Cravatt, B. F. Activity-based protein profiling: the serine hydrolases. *Proc. Natl Acad. Sci. USA* **96,** 14694–14699 (1999).
46. Patricelli, M. P., Giang, D. K., Stamp, L. M. & Burbaum, J. J. Direct visualization of serine hydrolase activities in complex proteome using fluorescent active site-directed probes. *Proteomics* **1,** 1067–1071 (2001).
47. Jessani, N. *et al.* A streamlined platform for high-content functional proteomics of primary human specimens. *Nature Methods* **2,** 691–697 (2005).
48. Leung, D., Hardouin, C., Boger, D. L. & Cravatt, B. F. Discovering potent and selective reversible inhibitors of enzymes in complex proteomes. *Nature Biotechnol.* **21,** 687–691 (2003).
49. Saghatelian, A. *et al.* Assignment of endogenous substrates to enzymes by global metabolite profiling. *Biochemistry* **43,** 14332–14339 (2004).
50. Chiang, K. P., Niessen, S., Saghatelian, A. & Cravatt, B. F. An enzyme that regulates ether lipid signaling pathways in cancer annotated by multidimensional profiling. *Chem. Biol.* **13,** 1041–1050 (2006).
51. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science* **298,** 1912–1934 (2002).
52. Matsuoka, S. *et al.* ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. *Science* **316,** 1160–1166 (2007).
53. Shiloh, Y. The ATM-mediated DNA-damage response: taking shape. *Trends Biochem. Sci.* **31,** 402–410 (2006).
54. Oda, Y., Huang, K., Cross, F. R., Cowburn, D. & Chait, B. T. Accurate quantitation of protein expression and site-specific phosphorylation. *Proc. Natl Acad. Sci. USA* **96,** 6591–6596 (1999).
55. Mann, M. Functional and quantitative proteomics using SILAC. *Nature Rev. Mol. Cell Biol.* **7,** 952–958 (2006).
56. Wang, B. *et al.* Abraxas and RAP80 form a BRCA1 protein complex required for the DNA damage response. *Science* **316,** 1194–1198 (2007).
57. Smogorzewska, A. *et al.* Identification of the FANCI protein, a monoubiquitinated FANCD2 paralog required for DNA repair. *Cell* **129,** 289–301 (2007).
58. Olsen, J. V. *et al.* Global, *in vivo*, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127,** 635–648 (2006).
59. Rush, J. *et al.* Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nature Biotechnol.* **23,** 94–101 (2005).
60. Jin, M. *et al.* Quantitative analysis of protein phosphorylation in mouse brain by hypothesis-driven multistage mass spectrometry. *Anal. Chem.* **77,** 7845–7851 (2005).
61. Huang, P. H. *et al.* Quantitative analysis of EGFRvIII cellular signaling networks reveals a combinatorial therapeutic strategy for glioblastoma. *Proc. Natl Acad. Sci. USA* **104,** 12867–12872 (2007).
62. Kim, S. C. *et al.* Substrate and functional diversity of lysine acetylation revealed by a proteomics survey. *Mol. Cell* **23,** 607–618 (2006).
63. Garcia, B. A., Pesavento, J. J., Mizzen, C. A. & Kelleher, N. L. Pervasive combinatorial modification of histone H3 in human cells. *Nature Methods* **4,** 487–489 (2007).
64. Ong, S. E., Mittler, G. & Mann, M. Identifying and quantifying *in vivo* methylation sites by heavy methyl SILAC. *Nature Methods* **1,** 119–126 (2004).
65. Khidekel, N. *et al.* Probing the dynamics of *O*-GlcNAc glycosylation in the brain using quantitative proteomics. *Nature Chem. Biol.* **3,** 339–348 (2007).
66. Peng, J. *et al.* A proteomics approach to understanding protein ubiquitination. *Nature Biotechnol.* **21,** 921–926 (2003).
67. Mukhopadhyay, A. & Tissenbaum, H. A. Reproduction and longevity: secrets revealed by *C. elegans*. *Trends Cell Biol.* **17,** 65–71 (2007).
68. Kenyon, C., Chang, J., Gensch, E., Rudner, A. & Tabtiang, R. A *C. elegans* mutant that lives twice as long as wild type. *Nature* **366,** 461–464 (1993).
69. Dong, M. Q. *et al.* Quantitative mass spectrometry identifies new insulin targets in *C. elegans*. *Science* **317,** 660–663 (2007).
70. Venable, J. D., Dong, M. Q., Wohlschlegel, J., Dillin, A. & Yates, J. R. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nature Methods* **1,** 39–45 (2004).
71. Liu, H., Sadygov, R. G. & Yates, J. R. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76,** 4193–4201 (2004).
72. Wu, C. C., MacCoss, M. J., Howell, K. E., Matthews, D. E. & Yates, J. R. Metabolic labeling of mammalian organisms with stable isotopes for quantitative proteomic analysis. *Anal. Chem.* **76,** 4951–4959 (2004).
73. Berns, K. *et al.* A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* **428,** 431–437 (2004).
74. Perrimon, N. & Mathey-Pervot, B. Applications of high-throughput RNA interference screens to problems in cell and developmental biology. *Genetics* **175,** 7–16 (2007).
75. Zabrouskov, V., Senko, M. W., Du, Y., Leduc, R. D. & Kelleher, N. L. New and automated MS^n approaches for top-down identification of modified proteins. *J. Am. Soc. Mass Spectrom.* **16,** 2027–2038 (2005).
76. Conrads, T. P., Anderson, G. A., Veenstra, T. D., Pasa-Tolic, L. & Smith, R. D. Utility of accurate mass tags for proteome-wide protein identification. *Anal. Chem.* **72,** 3349–3354 (2000).
77. Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W. & Gygi, S.P. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl Acad. Sci. USA* **100,** 6940–6945 (2003).
78. MacCoss, M. J. *et al.* Shotgun identification of protein modifications from protein complexes and lens tissue. *Proc. Natl Acad. Sci. USA* **99,** 7900–7905 (2002).

Acknowledgements We gratefully acknowledge the support of the National Institutes of Health.

Author information Reprints and permissions information is available at npg.nature.com/reprints. Correspondence should be addressed to B.F.C. (cravatt@scripps.edu) or J.R.Y. (jyates@scripps.edu).

Reaching for high-hanging fruit in drug discovery at protein–protein interfaces

James A. Wells[1,2] & Christopher L. McClendon[3]

Targeting the interfaces between proteins has huge therapeutic potential, but discovering small-molecule drugs that disrupt protein–protein interactions is an enormous challenge. Several recent success stories, however, indicate that protein–protein interfaces might be more tractable than has been thought. These studies discovered small molecules that bind with drug-like potencies to 'hotspots' on the contact surfaces involved in protein–protein interactions. Remarkably, these small molecules bind deeper within the contact surface of the target protein, and bind with much higher efficiencies, than do the contact atoms of the natural protein partner. Some of these small molecules are now making their way through clinical trials, so this high-hanging fruit might not be far out of reach.

There is probably no class of macromolecular interaction that rivals the complexity, diversity and regulatory impact of interactions between proteins[1–6]. There is much interest in targeting the interfaces between interacting proteins for therapeutic purposes — ideally with small 'drug-like' molecules, which generally are cheaper and can be administered orally instead of by injection — but the characteristics of these interfaces make this task a huge challenge. The contact surfaces involved in protein–protein interactions are large (~1,500–3,000 Å^2)[7,8] compared with those involved in protein–small-molecule interactions (~300–1,000 Å^2)[9,10]. In addition, the contact surfaces of proteins that interact with other proteins are generally flat and often lack the grooves and pockets present at the surfaces of proteins that bind to small molecules[11]. Unlike the classic proteins for which small-molecule drugs have been designed (for example, enzymes and G-protein-coupled receptors), protein–protein interactions do not have natural small-molecule partners. Thus, efforts to discover drugs that bind to a protein–protein interface do not have the luxury of starting from a small natural substrate or ligand.

Most contact surfaces in protein–protein interfaces also involve amino-acid residues that are not contiguous in the polymer chain. For this reason, peptides derived from short contiguous sequences at the interface are generally poor chemical starting points. (There are notable exceptions, however, in which a protein displays a contiguous tripeptide or tetrapeptide sequence for which small-molecule peptidomimetics have been assembled; see refs 12 and 13 for examples.) Furthermore, high-throughput screening (HTS) does not routinely identify compounds that disrupt protein–protein interfaces[14,15]. And, although biopharmaceuticals such as monoclonal antibodies and polypeptide hormones almost always bind to protein–protein interaction surfaces, there are few approved small-molecule drugs that do so.

Despite these challenges, several lines of evidence provide hope for finding small molecules that target protein–protein interfaces. Although these interfaces are large, mutational studies show that a small subset of the residues involved contributes most of the free energy of binding[16–20] (Fig. 1). Such 'hotspots' constitute less than half of the contact surface of a protein involved in the protein–protein interaction and are usually found at the centre of the contact interface. Proteins involved in protein–protein interactions can be 'promiscuous', binding to several targets using the same hotspot region[21]. Structural studies show that these promiscuous contact surfaces are adaptable, allowing one protein to engage a range of structurally diverse partners. Moreover, peptides selected for binding to one of the partners in a protein–protein pair (by using phage display) often compete with the natural protein partner for binding to the hotspot[20–24]. Thus, there seem to be many chemical solutions for tight binding, and large contact surfaces can be engaged by more-compact structures.

Research into finding small molecules that disrupt protein–protein interfaces has made considerable progress in the past five years (see refs 25–28 for recent reviews). Here, we focus on six recently published examples of discontinuous protein–protein interfaces for which small molecules that directly compete with one of the protein partners have been discovered (Fig. 2; Table 1). These examples are particularly instructive because crystal structures that are publicly available in the Protein Data Bank (PDB) allow comparison of the protein–protein complexes and the protein–small-molecule complexes. This provides the opportunity to analyse structurally how a small molecule directly competes with a natural protein partner. We also compare the affinities of the protein–protein complexes with those of the protein–small-molecule complexes, using binding equilibrium dissociation constants (K_d) or, in the absence of direct binding data, dissociation constants from competitive binding experiments (K_i) or half-maximal inhibitory concentrations (IC_{50}) from functional assays. We then use these examples to address common myths about targeting protein–protein interfaces. Together, these case studies uncover patterns that should help to advance drug discovery in this important field.

How small can we go?

The six examples of protein–protein interaction inhibitors that fit the criteria above are discussed in detail in this section.

IL-2 binders

The cytokine interleukin-2 (IL-2) has a key role in the activation of T cells and in the rejection of tissue grafts and is therefore of considerable medical interest. A series of small molecules that bind to IL-2 were produced at Sunesis Pharmaceuticals; the tightest binding of these

[1]Department of Pharmaceutical Chemistry, [2]Department of Cellular and Molecular Pharmacology, [3]Graduate Group in Biophysics, University of California at San Francisco, 1700 4th Street 503A, San Francisco, California 94156, USA.

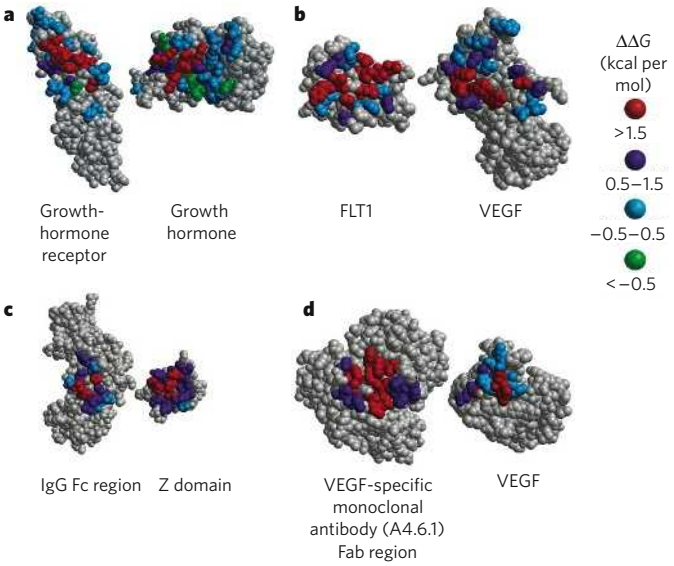


Figure 1 | Examples of protein–protein interface hotspots. Alanine-scanning mutational analysis (replacing each amino acid, in turn, with alanine) was carried out on the contact surfaces of four pairs of interacting proteins. The resultant change in the binding free energy compared with interactions involving the wild-type protein ($\Delta\Delta G$) is shown by colour coding amino-acid residues, from red (the most-disruptive changes) to green (those having little or no effect). Therefore, in each case, most of the free energy is contributed by a small number of residues (red), and this region is known as the hotspot[100]. VEGF, vascular endothelial growth factor; Z domain, a derivative of a domain from *Staphylococcus aureus* protein A. (Image courtesy of W. DeLano, DeLano Scientific, Palo Alto, California.)

molecules, SP4206 (Fig. 2; Table 1), has a dissociation constant in the mid-nanomolar range (K_i = 60 nM) and disrupts the interaction between IL-2 and the α-chain of the IL-2 receptor (IL-2Rα)[29–31]. These molecules were assembled in a fragment-based approach guided by X-ray structures and medicinal chemistry, and inspired by the previous drug-discovery efforts of Jefferson Tilley and co-workers at F. Hoffmann-La Roche[32]. Although the small molecules were assembled before the structure of the IL-2–IL-2Rα complex had been reported[33], they bind close to the centre of the receptor contact region on IL-2 (refs 34, 35) (Fig. 3a), and the interactions are specific because these molecules do not bind to IL-15, the closest homologue of IL-2 (ref. 31).

Several interesting features emerge when the structural and functional epitopes of IL-2 involved in binding to the small molecule SP4206 are compared with those involved in binding to the large protein IL-2Rα[19] (Fig. 3a; Supplementary Movie 1). The contact epitope for the small molecule is about half the size of that for the receptor. But, because the small molecule and the receptor bind to IL-2 with nearly equivalent affinities (less than a tenfold difference), the ligand efficiency (that is, the free energy of binding per non-hydrogen (heavy) atom[36]) for the small molecule is slightly less than twice that for the receptor (Table 1). The contact surface on IL-2 for binding to IL-2Rα is flat, as is typical of protein–protein complexes. By contrast, the small molecule traps a conformation of IL-2 in which a groove is present for small-molecule binding, and in which a loop of IL-2 has been repositioned to embrace the furanoic acid moiety at one end of the small molecule. Alanine-scanning mutational studies show that the small molecule and the receptor bind to the same hotspot residues on IL-2 (ref. 19). Although the structures of the small molecule and IL-2Rα differ markedly, the electrostatic potential of the surfaces presented is similar and probably reflects a need to establish electrostatic complementarity with IL-2. Electrostatic and surface-shape complementarity[37,38], as well as specific hydrogen-bonding interactions, probably account for the high selectivity of these interactions.

These studies show that the binding surface on IL-2 is adaptive and can bind to a small molecule with high affinity using the same main hotspot residues. It is notable that the design of this series of IL-2-binding small molecules did not require knowledge of the structure of the bound receptor complex. Instead, the design was informed by fragment-binding data and by structures of compounds bound to IL-2, coupled with medicinal chemistry and structure–activity relationships (SAR). Thus, the small molecule SP4206 is not an accurate atomic mimic of the receptor, and it would not have been discovered if it had been assumed that the precise structure of the receptor-bound form of IL-2 needed to be captured.

Bcl-X_L binders

Members of the B-cell lymphoma 2 (Bcl-2) family are important regulators of apoptotic cell death[39,40]. These molecules can form homodimers and can form heterodimers with other family members (generating various combinations of pro-apoptotic and/or anti-apoptotic subunits). For example, Bcl-2 and Bcl-X_L inhibit apoptosis by binding a 16-residue α-helical portion of the pro-apoptotic molecule BAK (Bcl-2-antagonist/killer)[40] or a 26-residue α-helical portion of another pro-apoptotic molecule, BAD (Bcl-2 antagonist of cell death)[41] (Fig. 3b). The importance of BAK and BAD as targets in the treatment of cancer has generated considerable interest in developing synthetic inhibitors of these protein–protein interactions. Several research groups have produced smaller helical molecules that mimic the key α-helix involved in this interaction and have high affinities ($K_i \approx$ 5–100 nM in the best cases)[42–45]. Recently, a team at Abbott Laboratories generated high-affinity organic compounds that bind to the hydrophobic helical domain of Bcl-X_L, Bcl-2 and another anti-apoptotic molecule, Bcl-W. These small molecules do not bind well to other anti-apoptotic members of the Bcl-2 family such as MCL1 (myeloid cell leukaemia sequence 1) and Bcl-B[46]. The most potent of these, ABT-737 (Fig. 2; Table 1), has a K_i of 0.6 nM and a molecular mass of 813 Da. Its affinity is therefore comparable to that of the α-helix, and because it has a smaller contact region with the protein, its ligand efficiency is almost twofold higher. The group of compounds was discovered using a fragment-based nuclear magnetic resonance (NMR) method known as SAR by NMR, and their properties were improved by using NMR-structure-guided medicinal chemistry[47–49]. The compounds were active in cell-based assays and in tumour xenograft models in animals, in which they showed synergy with several other chemotherapeutics and radiation. A derivative of ABT-737 — ABT-263 — is in phase I/II clinical trials for cancer (S. Fesik, personal communication).

NMR structures of small fragments that bind weakly to Bcl-X_L ($K_d \approx$ 0.3–4 mM) show ligand conformations similar to those of analogous groups in the elaborated high-affinity compounds these fragments gave rise to, such as ABT-737 (Supplementary Movie 2). Compared with the α-helix, however, there are marked differences (Fig. 3b). Alanine-scanning mutational analysis of the BAK-derived peptide identified several residues that are crucial for binding to Bcl-X_L: Val 74, Leu 78, Ile 81, Asp 83 and Ile 85 (refs 40, 45). The small molecule ABT-737 binds to the same region of Bcl-X_L as these residues of the BAK-derived peptide; however, it does not closely mimic the atomic details of the peptide. Instead, the small molecule traps a slightly different conformation of Bcl-X_L, binding in deeper cavities with more puckered grooves.

HDM2 binders

The human protein double minute 2 (HDM2) has emerged as an excellent drug target for cancer treatment. Initially, it was found that the mouse homologue of HDM2 (known as MDM2) binds to the tumour-suppressor protein p53 and increases its degradation, thus blocking the transcriptional activity of p53 that results in tumour suppression (see ref. 50 for a review). MDM2 can bind to a 15-residue α-helical region of p53 ($K_d \approx$ 600 nM), and the structure of the complex shows an interface that is largely hydrophobic[51]. Alanine-scanning mutational analysis of the 15-residue peptide identified three dominant amino acids in the centre of the interface: Phe 19, Trp 23 and Leu 26 (ref. 52). In search of inhibitors, a subsequent HTS and medicinal-chemistry effort at F. Hoffmann-La Roche (in Nutley, New Jersey) identified a series of tetra-substituted imidazoles, which the researchers named Nutlins. After considerable chemical optimization, the most potent of these small molecules, Nutlin-3 (Fig. 2; Table 1), was found to disrupt HDM2–p53 complexes with an IC_{50}

of 90 nM, and it showed potent p53-blocking activity *in vitro* and activity against tumour xenografts *in vivo*[53]. At Johnson & Johnson, 338,000 compounds were screened in parallel for binding to HDM2, by monitoring changes in thermostability (with a ThermoFluor instrument; Johnson & Johnson), and this resulted in the identification of a series of benzodiazepinediones[54]. After chemical optimization, one of these molecules (Fig. 2; Table 1) was found to have a high affinity for HDM2 (K_d = 67 nM, IC_{50} = 420 nM)[55]. Although these compounds were initially selected for binding to HDM2 and not for functional disruption of the complex, they promoted rapid dissociation of p53 from HDM2 in cells overproducing HDM2 (ref. 56). Furthermore, a benzodiazepinedione that had been modified with a solubilizing moiety inhibited the proliferation of tumour cells *in vitro* with an IC_{50} of ~10 μM and showed synergistic activity with the chemotherapeutic drug doxorubicin against tumours in mice[56].

The structures of the benzodiazepinedione bound to HDM2, and two Nutlins bound to the same protein, have been solved: HDM2–Nutlin-2 by X-ray crystallography[53], HDM2–Nutlin-3(analogue) by NMR spectroscopy[57], and HDM2 in complex with the benzodiazepinedione by X-ray crystallography[54] (Fig. 3c; Supplementary Movie 3). These three compounds bind to HDM2 within the same region as the α-helical portion of p53, and they insert aromatic or aliphatic moieties into hotspot pockets on HDM2 that bind to on HDM2 that bind to key residues on p53: Phe 19, Trp 23 and Leu 26. The contact epitopes for the small molecules are again about half the size of the minimal peptide-binding epitope. The conformation of HDM2 is more open at the ends when it binds to the peptide, whereas it closes more tightly over the small molecules, resulting in a more concave contact surface, as is the case for IL-2 and Bcl-X_L. It is remarkable that these dissimilar small-molecule scaffolds, discovered from markedly different starting points, have a similar mode of binding.

HPV E2 binders

Human papilloma virus (HPV) is of considerable interest, because it is the causative agent of warts and some cervical cancers. At present, there is no small-molecule drug that can treat these conditions. The interaction between the viral transcription factor E2 and the viral helicase E1 is crucial for the viral life cycle and thus is an important protein–protein-interface target. By using HTS, a research group at Boehringer Ingelheim identified a class of indandiones that moderately disrupts this interaction ($K_d \approx 20$ μM)[58]. Medicinal-chemistry efforts allowed further optimization of the affinity, with IC_{50} values as low as 6 nM[59–61] — for compound 23, for example (Fig. 2; Table 1). Direct binding of ^{3}H-labelled indandiones and isothermal titration calorimetry both showed that these small molecules bind to the transactivation domain of E2 with one-to-one stoichiometry. Interestingly, an X-ray structure of one of these small molecules, compound 18, bound to the E2 transactivation domain showed two copies of the small molecule; one penetrates into a cavity in the three-helix domain of E2, and the other sits on top[60].

Soon after the release of this X-ray structure, the structure of the transactivation domain of HPV type 18 (HPV-18) E2 in complex with E1 was reported[62]. The contact surface between E1 and E2 in HPV-18

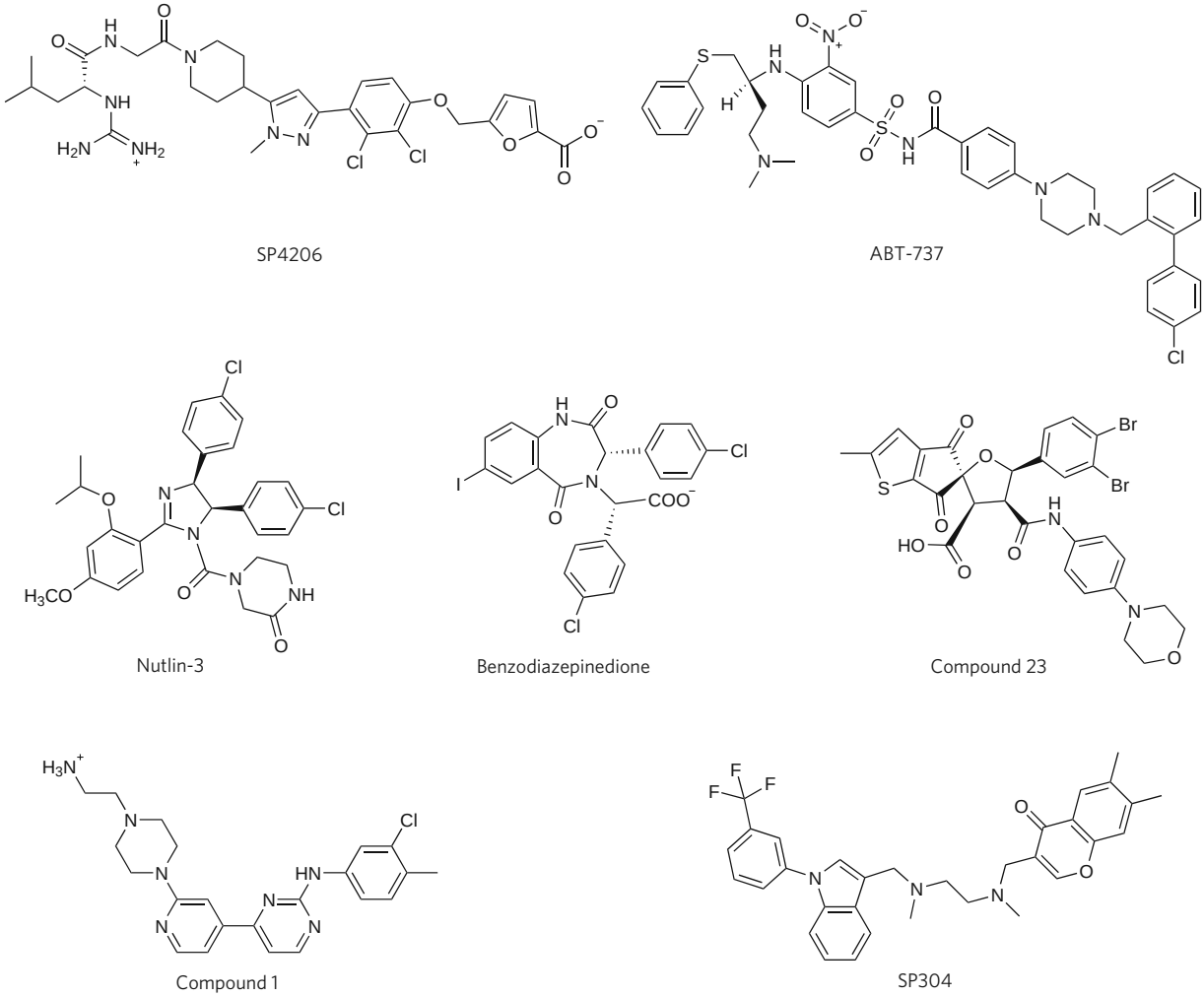


Figure 2 | Examples of small molecules that inhibit protein–protein interactions. Six examples of protein–protein interactions and the small molecules that have been discovered to inhibit them are described in this article. These compounds sit in different areas of chemical space from each other and most other pharmacophores. SP4206 binds to IL-2. ABT-737 binds to Bcl-X_L. Nutlin-3 and the benzodiazepinedione shown above bind to HDM2. Compound 23 binds to HPV E2. Compound 1 binds to ZipA. And SP304 binds to TNF. Selected physical and biochemical characteristics of these molecules are listed in Table 1.

Table 1 | Comparison of protein and small-molecule binding partners

Ligand	Molecular mass (Da)	PDB identity of complex	Affinity (μM)*	Ligand efficiency (kcal per mol per non-hydrogen atom)†	References
IL-2					
IL-2 receptor α-chain	24,790	1Z92	0.0105	0.11	33
SP4206	663	1PY2	0.06	0.21	31, 34
BCL-X_L					
BAD-derived peptide (amino acids 100–126)	3,110	2BZW	0.0006	0.16	89
ABT-737	813	2YXJ	0.0006	0.23	47
HDM2					
p53-derived peptide (amino acids 15–29)	1,808	1YCR	0.6	0.12	51
Nutlin-3	581	1RV1‡	0.09	0.24	53
Benzodiazepinedione	566	1T4E	0.067	0.31	54, 55
HPV E2					
E1	24,630	1TUE	0.06	0.14	62
Compound 23	684	1R6N‡	0.006	0.28	60, 61
ZipA					
FtsZ-derived peptide (amino acids 367–383)	2,024	1F47	21.6	0.13	63
Compound 1	425	1Y2F	12	0.23	67
TNF					
Subunit protein	17,381	1TNF	ND	ND	90
SP304	548	2AZ5	13	0.17	70

*Where a direct binding dissociation constant was not available or was not the lowest measured, the K_i or IC_{50} was used instead. †Ligand-efficiency values for the protein–protein pair are given as binding free energy ($-\Delta G$, in kcal per mol) per non-hydrogen contact atom, because so little of the protein is in contact. Ligand-efficiency values for the small molecule are given as binding free energy (kcal per mol) per non-hydrogen atom[36]. ‡The ligand in the X-ray structure is markedly similar to the compound listed. ND, not determined.

completely spans the E2–compound-18 contact interface in HPV-11 (Fig. 3d). Of the 20 residues of E2 that are in contact with E1, compound 18 is in contact with only 7. Importantly, compound 18 accesses a cavity that is not observed in the protein–protein interface (Supplementary Movie 4). Compound 23 achieves higher ligand efficiency than E1 (Table 1), presumably by deeply burying its hydrophobic surface area rather than spreading it across the interface.

ZipA binders

The separation of bacterial cells during cell division, and therefore their replication, depends on the formation of a septal ring. In certain Gram-negative bacteria, this ring consists of two or more proteins: FtsZ, a homologue of eukaryotic tubulin; and ZipA, a membrane-anchored protein. These molecules form a complex by interacting through their carboxy-terminal domains. A high-resolution (1.5 Å) X-ray structure revealed that a 17-residue peptide derived from the C terminus of *Escherichia coli* FtsZ binds to a cavity in ZipA as an extended β-strand followed by an α-helix[63] (Supplementary Movie 5). The FtsZ-derived peptide ($K_d \approx 20$ μM) binds about 100 times weaker than full-length FtsZ but is a useful surrogate. Although ten of the peptidyl side chains directly interact with ZipA, alanine-scanning mutational analysis shows that only four of these side chains (three hydrophobic and one acidic) dominate the binding affinity and constitute a hotspot. The structure of the unbound form of ZipA was also reported[63] and was found to be roughly similar to ZipA in complex with the FtsZ-derived peptide. A comparison of the pre-bound (apo) and the peptide-bound ZipA structures shows that loop adjustments and side-chain flips occur in ZipA to facilitate binding of the peptide. ZipA thus presents an adaptive surface for binding.

Using NMR spectroscopy to screen a diverse set of 825 compounds, Wyeth yielded 7 molecules that bound to ZipA at the same site as FtsZ[64]. Even though this is a high 'hit' rate (0.8%), indicating that ZipA might be 'druggable'[65], extensive medicinal-chemistry and SAR efforts starting from selected hits did not generate any high-affinity small molecules[66]. In a search for other possibilities, HTS of 250,000 compounds identified a pyridylpyrimidine, compound 1 (Fig. 2; Table 1), with a K_i of 12 μM[67]. The X-ray structure of compound 1 in complex with ZipA shows that the small molecule binds to ZipA entirely within the region bound by the 17-residue peptide and is in contact with only 740 $Å^2$ of ZipA compared with 1,350 $Å^2$ for the peptide (Supplementary Movie 5). Although the surface of this small molecule is more complementary to the surface of ZipA than the peptide surface is, these molecules could not penetrate deep into the surface of ZipA, unlike the other small molecules described here to bind IL-2, Bcl-X_L, HDM2 and HPV E2.

TNF disruptors

The cytokine tumour-necrosis factor (TNF) is a key factor in inflammatory responses and is therefore an important drug target. Biological therapeutics that target TNF have been approved for treating arthritis. Not surprisingly, there is considerable interest in developing small molecules or peptides that can disrupt the interaction between TNF and its receptors, TNFR1 and TNFR2. For example, small (13-residue) TNFR1-derived peptides that bind to TNF with moderate affinity ($K_d \approx 5$ μM) have been found[68], and photoactive small molecules that inhibit the TNF–TNFR1 interaction by labelling a site near where the receptor binds have also been discovered[69].

More recently, another class of small molecule that targets TNF was discovered, by using fragment screening[70]. These molecules disrupt TNF by binding (up to $K_d \approx 13$ μM) and displacing one of the three monomers that constitute TNF. More specifically, they bind to an adaptive cluster of tyrosine residues at the core of the trimer interface, which is illustrated in Fig. 4a for the small molecule SP304 (Fig. 2; Table 1). Two aromatic side chains from SP304 occupy the same position as the tyrosine residues from the displaced monomer (Supplementary Movie 6). Small molecules of this class are not seriously considered as drug candidates because of their moderate affinities; however, this finding shows that even constitutive interfaces in oligomeric proteins can bind to small molecules. Another recent example of this is the discovery of small molecules that can inhibit the activity of the anti-apoptotic protein survivin (also known as BIRC5) by binding at the interface between survivin homodimers[71].

It is known that TNF monomers in the trimer can be exchanged for free monomers, albeit slowly. Remarkably, the small molecules discussed above can increase the kinetics of monomer dissociation by more than 600-fold. Thus, instead of waiting for a monomer to dissociate completely (Fig. 4b, model 1), the small molecule can intercalate into the

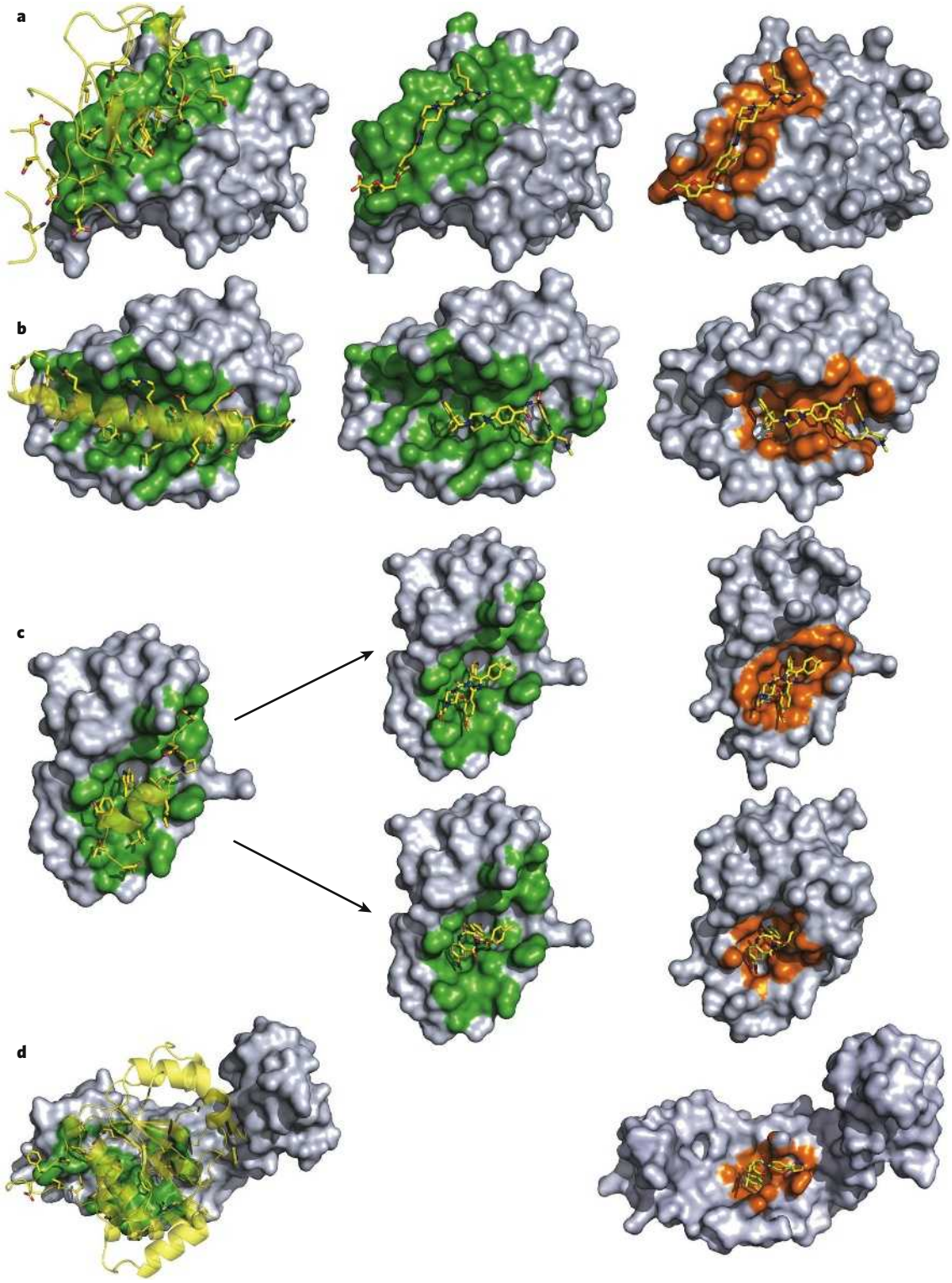


Figure 3 | Four comparisons of how a protein interacts with its natural protein (or peptide) partner and with a synthetic small molecule. The structures of protein–protein or protein–peptide complexes are shown on the left. The target protein is rendered as a filled surface (grey), and the binding protein or peptide is represented as a ribbon diagram (yellow), with selected side chains shown as sticks (with carbon in yellow, oxygen in red and nitrogen in blue). The contact surface on the target protein (within 4.5 Å of the binding partner) is shown in green. The structures of the protein–small-molecule complexes are shown on the right. The small molecule is shown in stick format, and the contact surface is shown in orange. In the centre, small molecules are shown superimposed on the protein in the conformation in which it binds to its natural protein or peptide partner, and the contact surface (on the target protein) of the natural interaction is shown in green. From these examples, it is clear that the protein–protein contact surface is much larger and flatter than the protein–small-molecule contact surface. **a**, IL-2 bound to its natural protein partner IL-2Rα (left), and IL-2 bound to the small molecule SP4206 (right). **b**, Bcl-X_L bound to a peptide derived from one of its natural protein partners, BAD, and Bcl-X_L bound to the small molecule ABT-737. **c**, HDM2 bound to a peptide derived from its natural protein partner p53, and HDM2 bound to the small molecule Nutlin-2 (upper) or a benzodiazepinedione (lower). **d**, HPV-18 E2 bound to HPV-18 E1, and HPV-11 E2 bound to the small molecule compound 18. The centre panel is not shown, because HPV-18 and HPV-11 are related but not identical. Images generated from files from the PDB, as indicated in Table 1.

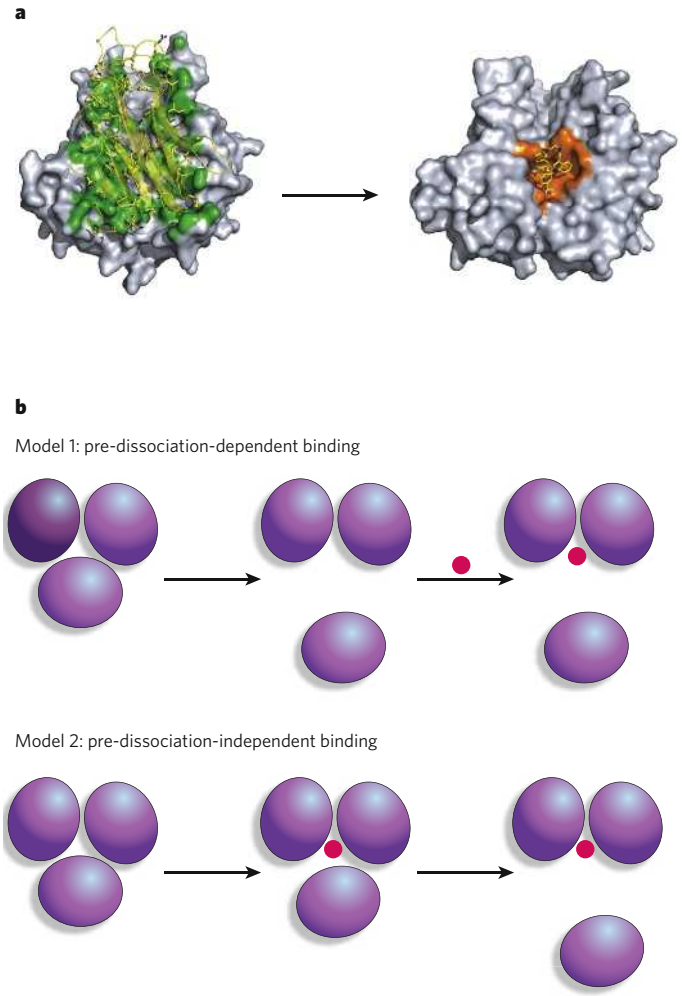


Figure 4 | Disruption of TNF by a small molecule. a, The structure of TNF, which is composed of three monomers, is shown on the left. Two of the TNF monomers are rendered as a filled surface (grey), and the other monomer is represented as a ribbon diagram (yellow). The contact surface on the TNF dimer (within 4.5 Å of the third monomer) is shown in green. The structure of the TNF dimer in complex with the small molecule SP304 is shown on the right. The small molecule is shown in stick format (with carbon in yellow, oxygen in red, nitrogen in blue and fluorine in white), and the contact surface on the TNF dimer is shown in orange. Images generated from files from the PDB, as indicated in Table 1. **b**, There are two models for how small molecules could block the formation of TNF trimers. In model 1, one of the monomers of TNF must completely dissociate before the small molecule can bind. In model 2, the small molecule can intercalate into the TNF complex and associate, which facilitates dissociation of a monomer. SP304 accelerates the rate of monomer dissociation (by more than 600-fold), which supports model 2.

dynamic trimer complex and displace the monomer (Fig. 4b, model 2). Presumably, the dynamic motions of proteins (see page 964) enable the small molecule to intercalate into the interface and prevent the displaced monomer from re-forming a high-affinity complex with the remaining dimer.

Myths about disrupting protein–protein interfaces

Until recently, lack of success in identifying small molecules that disrupt protein–protein interactions has led to several misconceptions about the prospects for new discoveries.

Protein–protein contact surfaces

One myth is that the large and flat contact surfaces seen in structures of protein complexes are rigid and do not present cavities for small molecules to bind. However, all of the contact surfaces described here show some adaptability, and cavities that are not seen in structures of either the free protein or the protein–protein complex are available for binding. Most of this flexibility involves motions of side chains and small perturbations of loops. In each case in Fig. 3, the small molecule accesses small pockets or grooves, which the larger, more constrained protein or peptide does not. Thus, it should not be assumed that the best binding site for a small molecule can be observed from static structures of either the free protein target or the protein–protein complex. For example, Bcl-X_L seems to have a rather flat surface in the static apo state, but during computer simulations of molecular dynamics in the absence of small molecules, transient pockets open in less than 1 nanosecond[72,73]. Similar transient openings of binding pockets were found in simulations with IL-2 and HDM2 (ref. 73).

Screening for protein–protein interface inhibitors

Another myth is that screening does not work for protein–protein interfaces. All of the examples presented here, however, involved empirical screening, either fragment screening or traditional HTS involving small-molecule libraries. In several examples, the starting compounds were identified by HTS, by using large numbers of compounds (more than 250,000) to identify moderate hits (K_i in the mid-micromolar range). Extensive biophysical techniques were applied to check that these hit compounds were 'real' and stoichiometric before any investment in medicinal-chemistry approaches. In four of the cases described here, medicinal-chemistry approaches improved the properties of these hit compounds to generate molecules with a K_d in the mid-nanomolar to low-nanomolar range; in two cases, ZipA and TNF, they did not. The ability to improve a hit compound was not well predicted by the behaviour of the initial compound or by inspection of the binding site, but it might be indicated by hit rates from fragment libraries[65] or by druggability indices[9] applied to possible protein–ligand conformations identified from computer simulations[72].

One explanation for why HTS is not more successful is that the libraries of compounds used for screening are derived mostly from historical medicinal-chemistry efforts in pharmaceutical companies. These 'chemical phenotypes' (chemotypes) are dominated by past drug-discovery research into G-protein-coupled receptors, enzymes and other traditional druggable targets. New classes of target often require new chemotypes. Thus, it is probable that each protein–protein interface will require a new chemotype. As a small-scale analysis, we took high-affinity inhibitors of protein–protein interactions by IL-2, Bcl-X_L, HDM2 and HPV E2, and we compared these small molecules with sets of compounds directed against targets in the chemical databases MDL Drug Data Report (MDDR; Symyx Technologies) and World of Molecular Bioactivity (WOMBAT; Sunset Molecular Discovery), by using a compound similarity ensemble approach[74]. The protein–protein interaction inhibitors did not show high similarity to any set of compounds against other known targets. Thus, if traditional libraries are used for screening, large collections of compounds might be required to find bona fide hit compounds with a K_d in the 10–100 μM range[75]. Moreover, it should not be assumed that there are a few 'privileged' scaffolds that will unlock this entire target class, as has been the case for protein kinases and G-protein-coupled receptors. Except for close homologues, each protein–protein interface is different, so the chemotypes of their inhibitors are likely to be more isolated in chemical space.

It is possible that fragment screening will be more successful than HTS when applied to protein–protein interfaces. Several successes have been achieved by using fragment screening, even though there have probably been far fewer fragment-screening efforts than HTS efforts. In theory, fragments (150–250 Da) have higher ligand efficiencies than typical compounds discovered by HTS (400–500 Da) and allow a greater search of chemical space[36,76].

Affinity of protein–protein interactions

A further myth is that native protein complexes have a higher affinity than protein–small-molecule complexes and cannot be competed away. In most of the cases described here, the optimized small molecule

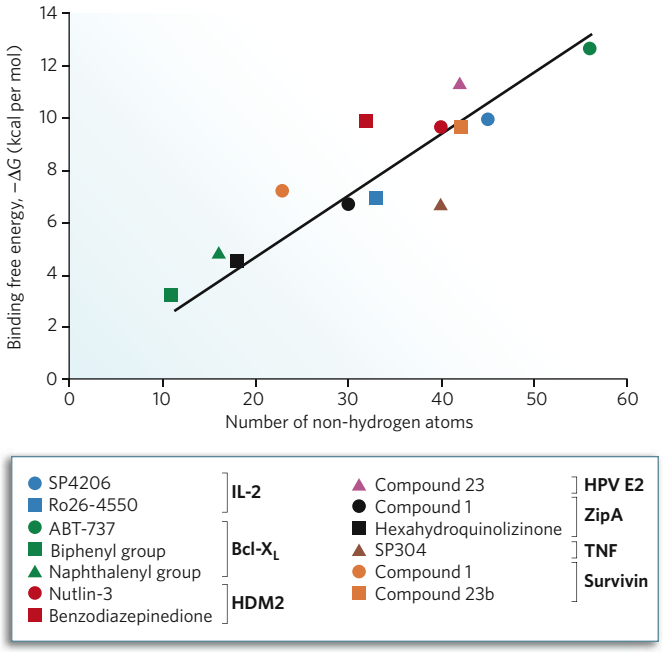


Figure 5 | Relationship between compound potency and size for small molecules that inhibit protein–protein interactions. For the highest-affinity fragments and small molecules that target protein–protein interfaces, the binding free energy ($-\Delta G$) is plotted against the number of non-hydrogen atoms. K_d values were converted to free energy (kcal per mol) using standard-state conditions of 1 M concentration at a temperature of 300 K. Where a direct binding dissociation constant was not available or was not the lowest measured, the K_i or IC_{50} was used instead. The slope can be described by $y = 0.24x$, and the correlation coefficient is 0.77. The linear relationship implies that there is a uniform ligand efficiency for these targets. Note that the first occurrence of Compound 1 (black circle) denotes the compound discussed in this article and depicted in Fig. 2, and the second occurrence (orange circle) denotes a molecule with a different chemical structure.

bound with an affinity comparable to that of the native partner protein or peptide. In several examples (IL-2, HDM2 and HPV E2), the K_i or IC_{50} was in the mid-nanomolar to low-nanomolar range and comparable to the binding affinity (K_d). This indicates that in equilibrium conditions, the small molecule is not at a disadvantage when it comes to displacing the protein partner.

From a kinetic perspective, small molecules might have an advantage over a large protein competitor, such as an antibody. For example, in the case of TNF, the inhibitory compounds accelerated the dissociation of a monomer from the TNF trimer by more than 600-fold. Thus, inhibition was not rate limited by the off-rate of a TNF monomer. It would be interesting to determine whether the other small molecules described here can accelerate dissociation of the respective protein–protein complexes. Recent paramagnetic NMR studies of protein complexes show that protein–protein interfaces might be inherently 'wobbly'[77,78]. If this is generally the case, then a small molecule could penetrate these dynamic 'encounter' complexes and have a kinetic advantage over a large antibody therapeutic, the association of which depends on complete (not partial) dissociation of the competing protein partner.

Size of small molecules that disrupt protein–protein interactions

Another myth is that small molecules that target protein–protein interfaces are too large to be drugs. For good oral absorption (or bioavailability), most orally active drugs are less than 500 Da[79,80], and drugs to treat neurological conditions usually need to have even lower molecular masses to cross the blood–brain barrier. Such criteria, derived from the limited set of known drugs, have notable exceptions; for example, cyclosporin A is ~1,000 Da. In addition, ABT-737 is 813 Da (Table 1) and has a reasonable (70%) bioavailability in rodents[46], and a derivative of comparable size, ABT-263, has entered clinical trials. Moreover, numerous drugs, including many antibiotics and anticancer drugs, are administered by injection, so in these cases, considerations of molecular mass are not driven by oral bioavailability.

There is always a trade-off between compound binding affinity and properties such as pharmacokinetics, solubility, toxicity and ease of synthesis, which together determine the probability that a compound will succeed as a drug. For optimum values of the latter properties, it is clearly better if the molecular mass is lower. All of the protein–protein interaction inhibitors described here with K_i values of less than 1 μM have molecular masses of 500–900 Da. We therefore wondered whether there is a limiting relationship between compound potency and size for small molecules that inhibit protein–protein interactions.

To analyse this, we selected only compounds for which there were extensive medicinal-chemistry data, as well as solved structures showing the compounds or close analogues bound to their targets. For the highest-affinity fragments and optimized compounds that bind to these target proteins, we plotted the binding free energy against the number of non-hydrogen atoms in the ligand (Fig. 5). These data have a reasonably linear distribution with a correlation coefficient of 0.77. It is remarkable that the small molecules that bind to these markedly different targets have similar ligand efficiencies, even though they belong to different chemotypes. The slope of the line gives a ligand efficiency of 0.24 kcal per mol per non-hydrogen atom. This value is considerably less than that for the tightest-binding small molecules such as biotin binding to avidin (1.15 kcal per mol per non-hydrogen atom)[36], but it is not

Table 2 | Ligand efficiencies of other small molecules that inhibit protein–protein interactions

Target	Compound	PDB identity of complex	Affinity (μM)	Ligand efficiency (kcal per mol per non-hydrogen atom)	References
Bcl-X_L	Compound 31	1YSI	0.036	0.27	49
HPV E2	Compound 18	1R6N	0.04	0.25	60, 61
ZipA	Compound 3	1Y2G	83.1	0.22	67
Clostridium botulinum neurotoxin B	Doxorubicin	1I1E	9.4	0.18	91
β-Catenin	PNU-74654	–	0.45	0.36	92
ARF1–ARNO complex	LM11	–	49.7	0.22	93
Dishevelled	FJ9	–	29	0.23	94
Rac	NSC23766	–	50	0.19	95
CD4 D1	J2	–	100	0.22	96
HIV gp120	NBD-556	–	47	0.26	97
EIF4E	4EGI-1	–	25	0.22	98
CD80	Compound 9	–	0.28	0.37	99

ARF1, ADP-ribosylation factor 1; ARNO, ARF nucleotide-binding-site opener (also known as PSCD2); D1, amino-terminal variable-region-like domain; EIF4E, eukaryotic translation initiation factor 4E; gp120, glycoprotein 120; HIV, human immunodeficiency virus.

dissimilar to that of many protein-kinase inhibitors (0.3–0.4 kcal per mol per non-hydrogen atom) and is comparable to that of many protease inhibitors (~0.25–0.35 kcal per mol per non-hydrogen atom)[36,81].

A survey of several less-optimized small molecules that inhibit protein–protein interfaces (Table 2) shows, with some exceptions, that these have ligand efficiencies similar to the small molecules listed in Table 1. Assuming a value of 0.24 kcal per mol per non-hydrogen atom, a compound with a K_d of 10 nM (typical of many drugs) would need to have ~46 non-hydrogen atoms (and therefore have a molecular mass of ~645 Da). We suggest that medicinal-chemistry efforts that generate molecules above this curve are doing exceptionally well. But for molecules that are considerably below this curve, much optimization is required if nanomolar affinity and oral bioavailability are desired.

Prospects and challenges for drug discovery

In the past five years, there has been remarkable progress in identifying, characterizing and developing small molecules that bind to protein–protein contact surfaces. In addition to binding to the contact surface itself, it is also possible to inhibit protein–protein interactions through allosteric sites[82,83] and by promoting aberrant interactions (for example, by cyclosporin A, which inhibits calcineurin by promoting an inhibitory interaction between cyclosporin A and cyclophilin A)[84]. However, there is still a long way to go before protein–protein interface inhibitors can be discovered routinely. It is not clear that the optimal region of compound space is being screened or that the compounds that are found can be easily optimized for these diverse interfaces. Fragment-screening methods offer excellent opportunities to cover a wider swathe of synthetically feasible chemical space per atom. The existence of hotspots means that ligand-efficient 'footholds' can be established by the initial fragments. However, except for the HDM2 binder, compounds that bound to the hotspot alone were not high-affinity inhibitors. To target IL-2, Bcl-X_L, HPV E2 and ZipA, additional sources of affinity were needed and were subsequently found (except for ZipA). The highest-affinity small molecules often engaged residues that the natural protein partner did not, exploiting 'cryptic' pockets within the protein contact interface.

For fragment screening to become more widely adopted, it will need to become cheaper, more sensitive and higher throughput. The biggest challenge in applying such fragment-screening technologies to drug discovery could be 'growing' fragments into higher-affinity small molecules. Improved computational methods to design ligand-efficient 'elaborated' compounds that bind to flexible protein sites would help to focus medicinal-chemistry efforts on these adaptive targets. In a recent study, high-affinity inhibitors of IL-2, Bcl-X_L and HDM2 were computationally docked to protein-conformation snapshots obtained from 10-nanosecond molecular-dynamics simulations[85]. In each case, one or more protein-conformation snapshots had a docked ligand conformation similar to the one observed experimentally. These results suggest that most, but not all, of the conformational differences seen when comparing unbound structures with inhibitor-bound structures result from conformational selection by the ligand. It is notable that the structural changes seen at these protein–protein interfaces are generally smaller than those that appear in classic examples of biologically evolved induced fit, such as hexokinase[86–88].

If we assume that protein–protein interactions have a lower 'ceiling' for ligand efficiency than more traditional targets, then the drug-discovery community will need to improve its management of the absorption, distribution, metabolism and excretion (ADME) properties of larger compounds. Although compounds that inhibit protein–protein interfaces are larger than typical drugs, these compounds are specific for their targets, as shown here for IL-2, Bcl-X_L, HDM2 and HPV E2. The compelling biology of protein–protein interfaces and the fact that several small molecules that inhibit protein–protein interactions are making their way through clinical trials provide hope that more of these drugs might be on the shelf in the future. Clearly, recent efforts have lifted us a rung higher in the quest to reach this class of high-hanging fruit. ■

1. Krogan, N. J. *et al.* Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440,** 637–643 (2006).
2. LaCount, D. J. *et al.* A protein interaction network of the malaria parasite *Plasmodium falciparum*. *Nature* **438,** 103–107 (2005).
3. Komurov, K. & White, M. Revealing static and dynamic modular architecture of the eukaryotic protein interaction network. *Mol. Syst. Biol.* **3,** 110 (2007).
4. Strong, M. & Eisenberg, D. The protein network as a tool for finding novel drug targets. *Prog. Drug Res.* **64,** 191–215 (2007).
5. Pu, S., Vlasblom, J., Emili, A., Greenblatt, J. & Wodak, S. J. Identifying functional modules in the physical interactome of *Saccharomyces cerevisiae*. *Proteomics* **7,** 944–960 (2007).
6. Collins, S. R. *et al.* Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* **446,** 806–810 (2007).
7. Jones, S. & Thornton, J. M. Principles of protein–protein interactions. *Proc. Natl Acad. Sci. USA* **93,** 13–20 (1996).
8. Conte, L. L., Chothia, C. & Janin, J. The atomic structure of protein–protein recognition sites. *J. Mol. Biol.* **285,** 2177–2198 (1999).
9. Cheng, A. C. *et al.* Structure-based maximal affinity model predicts small-molecule druggability. *Nature Biotechnol.* **25,** 71–75 (2007).
10. Smith, R. D. *et al.* Exploring protein–ligand recognition with Binding MOAD. *J. Mol. Graph. Model.* **24,** 414–425 (2006).
11. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nature Rev. Drug Discov.* **1,** 727–730 (2002).
12. Marsters, J. C. Jr *et al.* Benzodiazepine peptidomimetic inhibitors of farnesyltransferase. *Bioorg. Med. Chem.* **2,** 949–957 (1994).
13. Zobel, K. *et al.* Design, synthesis, and biological activity of a potent Smac mimetic that sensitizes cancer cells to apoptosis by antagonizing IAPs. *ACS Chem. Biol.* **1,** 525–533 (2006).
14. Robin, W. S. High-throughput screening of historic collections: observations on file size, biological targets, and file diversity. *Biotechnol. Bioeng.* **61,** 61–67 (1998).
15. Cochran, A. G. Antagonists of protein–protein interactions. *Chem. Biol.* **7,** R85–R94 (2000).
16. Clackson, T. & Wells, J. A. A hot spot of binding energy in a hormone-receptor interface. *Science* **267,** 383–386 (1995).
17. Clackson, T., Ultsch, M. H., Wells, J. A. & de Vos, A. M. Structural and functional analysis of the 1:1 growth hormone:receptor complex reveals the molecular basis for receptor affinity. *J. Mol. Biol.* **277,** 1111–1128 (1998).
18. Muller, Y. A. *et al.* Vascular endothelial growth factor: crystal structure and functional mapping of the kinase domain receptor binding site. *Proc. Natl Acad. Sci. USA* **94,** 7192–7197 (1997).
19. Thanos, C. D., DeLano, W. L. & Wells, J. A. Hot-spot mimicry of a cytokine receptor by a small molecule. *Proc. Natl Acad. Sci. USA* **103,** 15422–15427 (2006).
20. Moreira, I. S., Fernandes, P. A. & Ramos, M. J. Hot spots — a review of the protein–protein interface determinant amino-acid residues. *Proteins* **68,** 803–812 (2007).
21. DeLano, W. L., Ultsch, M. H., de Vos, A. M. & Wells, J. A. Convergent solutions to binding at a protein–protein interface. *Science* **287,** 1279–1283 (2000).
22. Sidhu, S. S., Lowman, H. B., Cunningham, B. C. & Wells, J. A. Phage display for selection of novel binding peptides. *Methods Enzymol.* **328,** 333–363 (2000).
23. Wrighton, N. C. *et al.* Small peptides as potent mimetics of the protein hormone erythropoietin. *Science* **273,** 458–464 (1996).
24. Livnah, O. *et al.* Functional mimicry of a protein hormone by a peptide agonist: the EPO receptor complex at 2.8 Å. *Science* **273,** 464–471 (1996).
25. Arkin, M. R. & Wells, J. A. Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. *Nature Rev. Drug Discov.* **3,** 301–317 (2004).
26. Yin, H. & Hamilton, A. D. Strategies for targeting protein–protein interactions with synthetic agents. *Angew. Chem. Int. Ed. Engl.* **44,** 4130–4163 (2005).
27. Fry, D. C. Protein–protein interactions as targets for small molecule drug discovery. *Biopolymers* **84,** 535–552 (2006).
28. Arkin, M. Protein–protein interactions and cancer: small molecules going in for the kill. *Curr. Opin. Chem. Biol.* **9,** 317–324 (2005).
29. Arkin, M. R. *et al.* Binding of small molecules to an adaptive protein–protein interface. *Proc. Natl Acad. Sci. USA* **100,** 1603–1608 (2003).
30. Braisted, A. C. *et al.* Discovery of a potent small molecule IL-2 inhibitor through fragment assembly. *J. Am. Chem. Soc.* **125,** 3714–3715 (2003).
31. Raimundo, B. C. *et al.* Integrating fragment assembly and biophysical methods in the chemical advancement of small-molecule antagonists of IL-2: an approach for inhibiting protein–protein interactions. *J. Med. Chem.* **47,** 3111–3130 (2004).
32. Tilley, J. W. *et al.* Identification of a small molecule inhibitor of the IL-2/IL-2Rα receptor interaction which binds to IL-2. *J. Am. Chem. Soc.* **119,** 7589–7590 (1997).
33. Rickert, M., Wang, X., Boulanger, M. J., Goriatcheva, N. & Garcia, K. C. The structure of interleukin-2 complexed with its α receptor. *Science* **308,** 1477–1480 (2005).
34. Thanos, C. D., Randal, M. & Wells, J. A. Potent small-molecule binding to a dynamic hot spot on IL-2. *J. Am. Chem. Soc.* **125,** 15280–15281 (2003).
35. Emerson, S. D. *et al.* NMR characterization of interleukin-2 in complexes with the IL-2Rα receptor component, and with low molecular weight compounds that inhibit the IL-2/IL-Rα interaction. *Protein Sci.* **12,** 811–822 (2003).
36. Kuntz, I. D., Chen, K., Sharp, K. A. & Kollman, P. A. The maximal affinity of ligands. *Proc. Natl Acad. Sci. USA* **96,** 9997–10002 (1999).
37. Lee, L. P. & Tidor, B. Optimization of binding electrostatics: charge complementarity in the barnase–barstar protein complex. *Protein Sci.* **10,** 362–377 (2001).
38. Midelfort, K. S. *et al.* Substantial energetic improvement with minimal structural perturbation in a high affinity mutant antibody. *J. Mol. Biol.* **343,** 685–701 (2004).
39. Adams, J. M. & Cory, S. The Bcl-2 protein family: arbiters of cell survival. *Science* **281,** 1322–1326 (1998).
40. Sattler, M. *et al.* Structure of Bcl-x_L–Bak peptide complex: recognition between regulators of apoptosis. *Science* **275,** 983–986 (1997).
41. Petros, A. M. *et al.* Rationale for Bcl-x_L/Bad peptide complex formation from structure, mutagenesis, and biophysical studies. *Protein Sci.* **9,** 2528–2534 (2000).
42. Sadowsky, J. D., Murray, J. K., Tomita, Y. & Gellman, S. H. Exploration of backbone space in foldamers containing α- and β-amino acid residues: developing protease-resistant

oligomers that bind tightly to the BH3-recognition cleft of Bcl-x_L. *Chembiochem* **8,** 903–916 (2007).
43. Yin, H. *et al.* Terphenyl-based Bak BH3 α-helical proteomimetics as low-molecular-weight antagonists of Bcl-x_L. *J. Am. Chem. Soc.* **127,** 10191-10196 (2005).
44. Walensky, L. D. *et al.* Activation of apoptosis *in vivo* by a hydrocarbon-stapled BH3 helix. *Science* **305,** 1466–1470 (2004).
45. Sadowsky, J. D. *et al.* (α/β+α)-peptide antagonists of BH3 domain/Bcl-x_L recognition: toward general strategies for foldamer-based inhibition of protein–protein interactions. *J. Am. Chem. Soc.* **129,** 139–154 (2007).
46. Oltersdorf, T. *et al.* An inhibitor of Bcl-2 family proteins induces regression of solid tumours. *Nature* **435,** 677–681 (2005).
47. Bruncko, M. *et al.* Studies leading to potent, dual inhibitors of Bcl-2 and Bcl-xL. *J. Med. Chem.* **50,** 641–662 (2007).
48. Hajduk, P. J. SAR by NMR: putting the pieces together. *Mol. Interv.* **6,** 266–272 (2006).
49. Petros, A. M. *et al.* Discovery of a potent inhibitor of the antiapoptotic protein Bcl-x_L from NMR and parallel synthesis. *J. Med. Chem.* **49,** 656–663 (2006).
50. Levine, A. J., Hu, W. & Feng, Z. The p53 pathway: what questions remain to be explored? *Cell Death Differ.* **13,** 1027–1036 (2006).
51. Kussie, P. H. *et al.* Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. *Science* **274,** 948–953 (1996).
52. Picksley, S. M., Vojtesek, B., Sparks, A. & Lane, D. P. Immunochemical analysis of the interaction of p53 with MDM2 — fine mapping of the MDM2 binding site on p53 using synthetic peptides. *Oncogene* **9,** 2523–2529 (1994).
53. Vassilev, L. T. *et al. In vivo* activation of the p53 pathway by small-molecule antagonists of MDM2. *Science* **303,** 844–848 (2004).
54. Grasberger, B. L. *et al.* Discovery and cocrystal structure of benzodiazepinedione HDM2 antagonists that activate p53 in cells. *J. Med. Chem.* **48,** 909–912 (2005).
55. Parks, D. J. *et al.* 1,4-Benzodiazepine-2,5-diones as small molecule antagonists of the HDM2–p53 interaction: discovery and SAR. *Bioorg. Med. Chem. Lett.* **15,** 765–770 (2005).
56. Koblish, H. K. *et al.* Benzodiazepinedione inhibitors of the Hdm2:p53 complex suppress human tumor cell proliferation *in vitro* and sensitize tumors to doxorubicin *in vivo*. *Mol. Cancer Ther.* **5,** 160–169 (2006).
57. Fry, D. C. *et al.* NMR structure of a complex between MDM2 and a small molecule inhibitor. *J. Biomol. NMR* **30,** 163–173 (2004).
58. Yoakim, C. *et al.* Discovery of the first series of inhibitors of human papillomavirus type 11: inhibition of the assembly of the E1–E2–Origin DNA complex. *Bioorg. Med. Chem. Lett.* **13,** 2539–2541 (2003).
59. White, P. W. *et al.* Inhibition of human papillomavirus DNA replication by small molecule antagonists of the E1–E2 protein interaction. *J. Biol. Chem.* **278,** 26765–26772 (2003).
60. Wang, Y. *et al.* Crystal structure of the E2 transactivation domain of human papillomavirus type 11 bound to a protein interaction inhibitor. *J. Biol. Chem.* **279,** 6976–6985 (2004).
61. Goudreau, N. *et al.* Optimization and determination of the absolute configuration of a series of potent inhibitors of human papillomavirus type-11 E1–E2 protein–protein interaction: a combined medicinal chemistry, NMR and computational chemistry approach. *Bioorg. Med. Chem.* **15,** 2690–2700 (2007).
62. Abbate, E. A., Berger, J. M. & Botchan, M. R. The X-ray structure of the papillomavirus helicase in complex with its molecular matchmaker E2. *Genes Dev.* **18,** 1981–1996 (2004).
63. Mosyak, L. *et al.* The bacterial cell-division protein ZipA and its interaction with an FtsZ fragment revealed by X-ray crystallography. *EMBO J.* **19,** 3179–3191 (2000).
64. Tsao, D. H. *et al.* Discovery of novel inhibitors of the ZipA/FtsZ complex by NMR fragment screening coupled with structure-based design. *Bioorg. Med. Chem.* **14,** 7953–7961 (2006).
65. Hajduk, P. J., Huth, J. R. & Fesik, S. W. Druggability indices for protein targets derived from NMR-based screening data. *J. Med. Chem.* **48,** 2518–2525 (2005).
66. Jennings, L. D. *et al.* Combinatorial synthesis of substituted 3-(2-indolyl)piperidines and 2-phenyl indoles as inhibitors of ZipA–FtsZ interaction. *Bioorg. Med. Chem.* **12,** 5115–5131 (2004).
67. Rush, T. S., Grant, J. A., Mosyak, L. & Nicholls, A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein–protein interaction. *J. Med. Chem.* **48,** 1489–1495 (2005).
68. Takasaki, W., Kajino, Y., Kajino, K., Murali, R. & Greene, M. I. Structure-based design and characterization of exocyclic peptidomimetics that inhibit TNFα binding to its receptor. *Nature Biotechnol.* **15,** 1266–1270 (1997).
69. Carter, P. H. *et al.* Photochemically enhanced binding of small molecules to the tumor necrosis factor receptor-1 inhibits the binding of TNF-α. *Proc. Natl Acad. Sci. USA* **98,** 11879–11884 (2001).
70. He, M. M. *et al.* Small-molecule inhibition of TNF-α. *Science* **310,** 1022–1025 (2005).
71. Wendt, M. D. *et al.* Discovery of a novel small molecule binding site of human survivin. *Bioorg. Med. Chem. Lett.* **17,** 3122–3129 (2007).
72. Brown, S. P. & Hajduk, P. J. Effects of conformational dynamics on predicted protein druggability. *ChemMedChem* **1,** 70–72 (2006).
73. Eyrisch, S. & Helms, V. Transient pockets on protein surfaces involved in protein–protein interaction. *J. Med. Chem.* **50,** 3457–3464 (2007).
74. Keiser, M. J. *et al.* Relating protein pharmacology by ligand chemistry. *Nature Biotechnol.* **25,** 197–206 (2007).
75. Feng, B. Y. *et al.* A high-throughput screen for aggregation-based inhibition in a large compound library. *J. Med. Chem.* **50,** 2385–2390 (2007).
76. Carr, R. A. E., Congreve, M., Murray, C. W. & Rees, D. C. Fragment-based lead discovery: leads by design. *Drug Discov. Today* **10,** 987–992 (2005).
77. Volkov, A. N., Worrall, J. A., Holtzmann, E. & Ubbink, M. Solution structure and dynamics of the complex between cytochrome *c* and cytochrome *c* peroxidase determined by paramagnetic NMR. *Proc. Natl Acad. Sci. USA* **103,** 18945–18950 (2006).
78. Tang, C., Iwahara, J. & Clore, G. M. Visualization of transient encounter complexes in protein–protein association. *Nature* **444,** 383–386 (2006).
79. Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **44,** 235–249 (2000).
80. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **46,** 3–26 (2001).
81. Erlanson, D. A. Fragment-based lead discovery: a chemical update. *Curr. Opin. Biotechnol.* **17,** 643–652 (2006).
82. Lowe, J., Li, H., Downing, K. H. & Nogales, E. Refined structure of αβ-tubulin at 3.5 Å resolution. *J. Mol. Biol.* **313,** 1045–1057 (2001).
83. McMillan, K. *et al.* Allosteric inhibitors of inducible nitric oxide synthase dimerization discovered via combinatorial chemistry. *Proc. Natl Acad. Sci USA* **97,** 1506–1511 (2000).
84. Ke, H. & Huai, Q. Crystal structures of cyclophilin and its partners. *Front. Biosci.* **9,** 2285–2296 (2004).
85. Eyrisch, S. & Helms, V. Transient pockets on protein surfaces involved in protein–protein interaction. *J. Med. Chem.* **50,** 3457–3464 (2007).
86. Fletterick, R. J., Bates, D. J. & Steitz, T. A. The structure of a yeast hexokinase monomer and its complexes with substrates at 2.7-Å resolution. *Proc. Natl Acad. Sci. USA* **72,** 38–42 (1975).
87. Anderson, C. M., Zucker, F. H. & Steitz, T. A. Space-filling models of kinase clefts and conformation changes. *Science* **204,** 375–380 (1979).
88. Yankeelov, J. A. & Koshland, D. E. Evidence for conformation changes induced by substrates of phosphoglucomutase. *J. Biol. Chem.* **240,** 1593–1602 (1965).
89. Kelekar, A., Chang, B. S., Harlan, J. E., Fesik, S. W. & Thompson, C. B. Bad is a BH3 domain-containing protein that forms an inactivating dimer with Bcl-X_L. *Mol. Cell. Biol.* **17,** 7040–7046 (1997).
90. Eck, M. J. & Sprang, S. R. The structure of tumor necrosis factor-α at 2.6 Å resolution. Implications for receptor binding. *J. Biol. Chem.* **264,** 17595–17605 (1989).
91. Eswaramoorthy, S., Kumaran, D. & Swaminathan, S. Crystallographic evidence for doxorubicin binding to the receptor-binding site in *Clostridium botulinum* neurotoxin B. *Acta Crystallogr. D Biol. Crystallogr.* **57,** 1743–1746 (2001).
92. Trosset, J.-Y. *et al.* Inhibition of protein–protein interactions: the discovery of druglike β-catenin inhibitors by combining virtual and biophysical screening. *Proteins* **64,** 60–67 (2006).
93. Viaud, J. *et al.* Structure-based discovery of an inhibitor of Arf activation by Sec7 domains through targeting of protein–protein complexes. *Proc. Natl Acad. Sci. USA* **104,** 10370–10375 (2007).
94. Fujii, N. *et al.* An antagonist of dishevelled protein–protein interaction suppresses β-catenin-dependent tumor cell growth. *Cancer Res.* **67,** 573–579 (2007).
95. Gao, Y., Dickerson, J. B., Guo, F., Zheng, J. & Zheng, Y. Rational design and characterization of a Rac GTPase-specific small molecule inhibitor. *Proc. Natl Acad. Sci. USA* **101,** 7618–7623 (2004).
96. Xiao, H. *et al.* Potent inhibition of the CD4-dependent T cell response by J2, a novel nonpeptide organic ligand of CD4 D1. *Mol. Immunol.* **44,** 784–795 (2007).
97. Schon, A. *et al.* Thermodynamics of binding of a low-molecular-weight CD4 mimetic to HIV-1 gp120. *Biochemistry* **45,** 10973–10980 (2006).
98. Moerke, N. J. *et al.* Small-molecule inhibition of the interaction between the translation initiation factors eIF4E and eIF4G. *Cell* **128,** 257–267 (2007).
99. Uvebrant, K. *et al.* Discovery of selective small-molecule CD80 inhibitors. *J. Biomol. Screen.* **12,** 464–472 (2007).
100. DeLano, W. L. Unraveling hot spots in binding interfaces: progress and challenges. *Curr. Opin. Struct. Biol.* **12,** 14–20 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature

Acknowledgements We thank M. Keiser for carrying out the similarity ensemble approach analysis, W. DeLano for MacPyMOL movie advice and for providing Fig. 1, M. Arkin and J. Sadowsky for proof-reading the manuscript, and M. Jacobson and B. Shoichet for discussions.

Author information Reprints and permissions information is available at npg.nature.com/reprints. Correspondence should be addressed to J.A.W. (jim.wells@ucsf.edu).

Kimberlite ascent and eruption

Arising from: L. Wilson & J. W. Head III *Nature* 447, 53–57 (2007).

Wilson and Head[1] model kimberlite ascent and eruption by considering the propagation of a volatile-rich dyke. Wilson and Head's model has features in common with Sparks *et al.*[2], but it is inconsistent with geological observations and constraints on volatile solubility. Here we show that this may be due to erroneous physical assumptions.

Dyke propagation is dependent on balances between buoyancy, source pressure and fracture strength[3,4]. Wilson and Head assume that kimberlite dykes are connected to the deep source and that the pressure gradient between the source and the dyke tip is governed by the release of copious carbon dioxide (CO_2). Thus, assumptions are made about the volume of available magma, CO_2 solubility and volatile composition, as well as about whether source pressure or buoyancy is dominant and about the behaviour of volatiles released into the crack tip. Wilson and Head state that 90% of the CO_2 is exsolved at 2 GPa. However, CO_2 becomes increasingly soluble as melts become more silica-deficient[5]; at 100 MPa, silica-poor basic melts can dissolve >1% CO_2 and, with a linear solubility law, most if not all CO_2 would be dissolved at 2 GPa. Furthermore, in carbonate-rich melts, most carbon is speciated as carbonate rather than molecular CO_2, as indicated by magmatic calcite in hypabyssal kimberlites[6]. The Wilson and Head model overestimates the amount of volatiles available to act as an exsolving propellant. Water may be a major volatile in kimberlite[2], but it only exsolves at low pressure.

In the model of Wilson and Head, volatiles are released from exsolving magma into the dyke tip with a very low pressure, resulting in very high pressure gradients and very high propagation speeds (tens of metres per second). However, experimental and theoretical studies[4,7] show that the much larger buoyancy of released volatiles results in a fluid-filled fracture accelerating in advance of the magma-filled dyke, consistent with observations from kimberlite dykes[8]. The pressure in the volatile-filled fracture moving in advance of and accelerating away from the magma must be at least the lithostatic pressure plus the mantle fracture strength, so we question the very low pressures, except for a negligibly small region at the volatile-filled crack tip[3,4]. Wilson and Head infer a decelerating fracture system, whereas previous work[9] on dyke nucleation indicates that acceleration is a consequence of the increase in length as dykes propagate and decompress.

There are difficulties reconciling the very short eruption times estimated by Wilson and Head and the geological complexity of kimberlites[2] (C. R. Clement *et al.*, unpublished results), which indicate prolonged multistage eruptions. Furthermore, constraints on volumes and magma supply rates through established dyke systems[2] indicate eruption times of days to months rather than an hour. Wilson and Head estimate large adiabatic coolings, but these are not consistent with estimates of high emplacement temperatures (>400 °C to 1,100 °C) of kimberlitic pyroclastics and hypabyssal intrusions[2,10,11].

The pipe-formation process proposed by Wilson and Head is unclear, but we envisage that it involves the principles of rock mechanics[2,12], combined with large early overpressures and later underpressures associated with explosive flows[2]. The geology supports a progressive, multistage and long-lived failure of wall-rocks by a variety of failure mechanisms rather than catastrophic pipe formation[2,12]. The fluidization wave model of Wilson and Head is evidently a dynamic phenomenon. Fluidization is usually applied in geological systems using concepts from engineering[13,14], in which gas flows continuously through unconsolidated granular materials. There is geological and experimental evidence that fluidization occurred in the waning pipe-filling stage of kimberlite eruptions[2,13,14].

We agree with Wilson and Head that fast transport aids diamond preservation, but there are other important factors because kimberlites contain mixtures of perfectly shaped, broken and resorbed diamonds[15], indicating diverse interaction histories with kimberlite magmas. Diamonds can be preserved within nodules, preventing reaction with kimberlite, and are released progressively during ascent by fragmentation of xenoliths, resulting in a range of interaction times[15].

R. S. J. Sparks[1], R. J. Brown[1], M. Field[1,2] & M. Gilbertson[3]
[1]Department of Earth Sciences, University of Bristol, Bristol BS8 1RJ, UK.
e-mail: steve.sparks@bristol.ac.uk
[2]De Beers MRM Group, Wells, Somerset BA5 3DG, UK.
[3]Department of Mechanical Engineering, University of Bristol, Bristol BS8 1TR, UK.

Received 26 July 2007; accepted 16 October 2007.

1. Wilson, L. & Head, J. W. III. An integrated model of kimberlite ascent and eruption. *Nature* **447,** 53–57 (2007).
2. Sparks, R. S. J. *et al.* Dynamics of kimberlite volcanism. *J. Volcanol. Geotherm. Res.* **155,** 18–48 (2006).
3. Lister, J. R. & Kerr, R. C. Fluid-mechanical models of crack propagation and their application to magma transport in dykes. *J. Geophys. Res.* **96,** 10049–10077 (1991).
4. Menand, T. & Tait, S. R. The propagation of a buoyant liquid-filled fissure from a source under constant pressure: an experimental approach. *J. Geophys. Res.* 107, 2306 16–1-14 (2002).
5. Brooker, R. A., Kohn, S., Holloway, J. R. & McMillan, P. F. Structural controls on the solubility of CO_2 in silicate melts. Part I: bulk solubility data. *Chem. Geol.* **174,** 225–239 (2001).
6. Mitchell, R. H. *Kimberlites: Mineralogy, Geochemistry and Petrology* (Plenum, New York, 1986).
7. Menand, T. & Tait, S. R. A phenomenological model for precursor volcanic eruptions. *Nature* **411,** 678–680 (2001).
8. Brown, R. J., Kavanagh, J., Sparks, R. S. J., Tait, M. & Field, M. Mechanically disrupted and chemically weakened zones in segmented kimberlite dike systems cause the localisation of kimberlites. *Geology* **35,** 815–818 (2007).
9. McLeod, P. & Tait, S. R. The growth of dykes from magma chambers. *J. Volcanol. Geotherm. Res.* **92,** 231–245 (1999).
10. Fedortchouk, Y. & Canil, D. Intensive variables in kimberlite magmas, Lac de Gras, Canada and implications for diamond survival. *J. Petrol.* **45,** 1725–1745 (2004).
11. Stripp, G., Field, M., Schumacher, J. C. & Sparks, R. S. J. Post-emplacement serpentinisation and related hydrothermal metamorphism in a kimberlite from Venetia, South Africa. *J. Metamorph. Geol.* **24,** 515–534 (2006).
12. Barnett, W. The rock mechanics of kimberlite pipe formation. *J. Volcanol. Geotherm. Res.* (in the press).
13. Walters, A. L. *et al.* The role of fluidisation in the formation of volcaniclastic kimberlite: grain size observations and experimental investigation. *J. Volcanol. Geotherm. Res.* **155,** 119–137 (2006).
14. Gernon, T., Gilbertson, M. A., Sparks, R. S. J. & Field, M. Gas-fluidisation in an experimental tapered bed: insights into processes in diverging volcanic conduits. *J. Volcanol. Geotherm. Res.* (in the press).
15. Gurney, J. in *Kimberlites and related rocks: their mantle/crust setting, diamonds and diamond exploration. Proceedings of the Fourth International Kimberlite Conference.* (eds Ross, J. *et al.*) *Geol. Soc. Australia Special Publication*, **14,** 935–965, 990–1000 (Perth, Australia, 1989).

doi:10.1038/nature06435

Wilson & Head reply

Replying to: R. S. J. Sparks, R. J. Brown, M. Field & M. Gilbertson *Nature* **450,** doi:10.1038/nature06435 (2007).

Differences between the model of Sparks *et al.*[1,2] and ours[3] arise mainly because we focus on phenomena during the transient, opening phase that we suggest dominates many kimberlite eruptions, rather than on the subsequent, more prolonged phases relevant to other kimberlite eruptions[1,2].

If water dominates carbon dioxide (CO_2) as the vapour phase[1,2], our argument[3] about the pressure distribution driving a kimberlite dyke to the surface is reinforced. The key factor allowing the initial rapid ascent is the large difference between the high mantle source pressure and the low dyke tip pressure, the latter being buffered by the saturation pressure of the least soluble volatile phase[4,5]. The dyke tip pressure required for water to exsolve will be even lower than the pressure we inferred for CO_2, thus increasing the pressure difference driving the magma upward through the opening dyke.

Any vapour-filled region at the tip of a dyke, breaking away to propagate faster as an independent crack, can incorporate[6] some of the magmatic foam implied by our model[3]. Cracks longer than ~20 m travel at ~1 km s^{-1}, which is ~40% of the sound speed in rock[7]. Chilling of magma in the closing crack base 'heals' the fracture, restoring the country rock mechanical properties. When the dyke tip subsequently arrives, it encounters essentially the same conditions as if crack separation had never occurred. Only seconds are needed to chill a 1–2-mm-thick film of magma left behind by a ~20-m-long crack; during this time the dyke tip, rising[3] at ~20 m s^{-1}, travels ~100 m—a tiny fraction of the dyke's vertical extent. A new low-pressure region starts to grow below the dyke tip immediately after crack separation; we infer that the stress and pressure conditions we proposed will be present over most of the path of the rising dyke tip.

Our dyke geometries are only slightly larger than those of Sparks *et al.*[1,2], and their minimum estimates of total magma volumes imply eruption durations only a few times longer than the time to establish the dyke pathway[3]; larger volumes will imply more prolonged events. Our calculations[3] of adiabatic cooling refer to magma reaching the surface during the opening phase of an eruption; in a long-lived eruption, most of the magma finally emplaced in the sub-surface diatreme will indeed suffer less cooling.

We suggested[3] a violent change from overpressure to underpressure as a dyke reached the surface, with rapid physical development of the near-surface pipe and diatreme system. A longer-lived eruption[1,2] will indeed allow a range of additional failure mechanisms. Although 'fluidization' commonly relates to the near-steady passage of gas through unconsolidated granular materials, as in the waning phases of kimberlite eruptions[8], the basic physics is the same as that in our violent opening phase.

Regarding preservation of diamonds during transit to the surface, we stress[3] that rapid transport will maximise the survival of diamonds as they pass through potentially unstable combinations of ambient pressure and temperature conditions, irrespective of the chemical environment that they encounter[9].

Lionel Wilson[1] & James W. Head III[2]
[1]Environmental Science Department, Lancaster University, Lancaster LA1 4YQ, UK.
[2]Geological Sciences Department, Brown University, Providence, Rhode Island 02912, USA.
e-mail: james_head@brown.edu

1. Sparks, R. S. J. *et al.* Dynamics of kimberlite volcanism. *J. Volcanol. Geotherm. Res.* **155,** 18–48 (2006).
2. Sparks, R. S. J., Brown, R. J., Field, M. & Gilbertson, M. Kimberlite ascent and eruption. *Nature* **450,** doi:10.1038/nature06435 (2007).
3. Wilson, L. & Head, J. W. III. An integrated model of kimberlite ascent and eruption. *Nature* **447,** 53–57 (2007).
4. Lister, J. R. & Kerr, R. C. Fluid-mechanical models of crack propagation and their application to magma transport in dykes. *J. Geophys. Res.* **96,** 10049–10077 (1991).
5. Rubin, A. M. Dikes vs. diapirs in viscoelastic rock. *Earth Planet. Sci. Lett.* **119,** 641–659 (1993).
6. Menand, T. & Tait, S. R. A phenomenological model for precursor volcanic eruptions. *Nature* **411,** 678–680 (2001).
7. Dobran, F. *Volcanic Processes — Mechanisms in Material Transport* page 212 (Kluwer/Plenum, New York, 2001).
8. Walters, A. L. *et al.* The role of fluidisation in the formation of volcaniclastic kimberlite: grain size observations and experimental investigation. *J. Volcanol. Geotherm. Res.* **155,** 119–137 (2006).
9. Ross, J. *et al.* (eds) *Kimberlites and related rocks: their mantle/crust setting, diamonds and diamond exploration. Proceedings of the Fourth International Kimberlite Conference.* Geol. Soc. Australia Special Publication, **14,** 935–965, 990–1000 (Perth, Australia, 1989).

doi:10.1038/nature06436

Kimberlite ascent and eruption

Arising from: L. Wilson & J. W. Head III *Nature* 447, 53–57 (2007).

Wilson and Head[1] model kimberlite ascent and eruption by considering the propagation of a volatile-rich dyke. Wilson and Head's model has features in common with Sparks *et al.*[2], but it is inconsistent with geological observations and constraints on volatile solubility. Here we show that this may be due to erroneous physical assumptions.

Dyke propagation is dependent on balances between buoyancy, source pressure and fracture strength[3,4]. Wilson and Head assume that kimberlite dykes are connected to the deep source and that the pressure gradient between the source and the dyke tip is governed by the release of copious carbon dioxide (CO_2). Thus, assumptions are made about the volume of available magma, CO_2 solubility and volatile composition, as well as about whether source pressure or buoyancy is dominant and about the behaviour of volatiles released into the crack tip. Wilson and Head state that 90% of the CO_2 is exsolved at 2 GPa. However, CO_2 becomes increasingly soluble as melts become more silica-deficient[5]; at 100 MPa, silica-poor basic melts can dissolve >1% CO_2 and, with a linear solubility law, most if not all CO_2 would be dissolved at 2 GPa. Furthermore, in carbonate-rich melts, most carbon is speciated as carbonate rather than molecular CO_2, as indicated by magmatic calcite in hypabyssal kimberlites[6]. The Wilson and Head model overestimates the amount of volatiles available to act as an exsolving propellant. Water may be a major volatile in kimberlite[2], but it only exsolves at low pressure.

In the model of Wilson and Head, volatiles are released from exsolving magma into the dyke tip with a very low pressure, resulting in very high pressure gradients and very high propagation speeds (tens of metres per second). However, experimental and theoretical studies[4,7] show that the much larger buoyancy of released volatiles results in a fluid-filled fracture accelerating in advance of the magma-filled dyke, consistent with observations from kimberlite dykes[8]. The pressure in the volatile-filled fracture moving in advance of and accelerating away from the magma must be at least the lithostatic pressure plus the mantle fracture strength, so we question the very low pressures, except for a negligibly small region at the volatile-filled crack tip[3,4]. Wilson and Head infer a decelerating fracture system, whereas previous work[9] on dyke nucleation indicates that acceleration is a consequence of the increase in length as dykes propagate and decompress.

There are difficulties reconciling the very short eruption times estimated by Wilson and Head and the geological complexity of kimberlites[2] (C. R. Clement *et al.*, unpublished results), which indicate prolonged multistage eruptions. Furthermore, constraints on volumes and magma supply rates through established dyke systems[2] indicate eruption times of days to months rather than an hour. Wilson and Head estimate large adiabatic coolings, but these are not consistent with estimates of high emplacement temperatures (>400 °C to 1,100 °C) of kimberlitic pyroclastics and hypabyssal intrusions[2,10,11].

The pipe-formation process proposed by Wilson and Head is unclear, but we envisage that it involves the principles of rock mechanics[2,12], combined with large early overpressures and later underpressures associated with explosive flows[2]. The geology supports a progressive, multistage and long-lived failure of wall-rocks by a variety of failure mechanisms rather than catastrophic pipe formation[2,12]. The fluidization wave model of Wilson and Head is evidently a dynamic phenomenon. Fluidization is usually applied in geological systems using concepts from engineering[13,14], in which gas flows continuously through unconsolidated granular materials. There is geological and experimental evidence that fluidization occurred in the waning pipe-filling stage of kimberlite eruptions[2,13,14].

We agree with Wilson and Head that fast transport aids diamond preservation, but there are other important factors because kimberlites contain mixtures of perfectly shaped, broken and resorbed diamonds[15], indicating diverse interaction histories with kimberlite magmas. Diamonds can be preserved within nodules, preventing reaction with kimberlite, and are released progressively during ascent by fragmentation of xenoliths, resulting in a range of interaction times[15].

R. S. J. Sparks[1], R. J. Brown[1], M. Field[1,2] & M. Gilbertson[3]
[1]Department of Earth Sciences, University of Bristol, Bristol BS8 1RJ, UK.
e-mail: steve.sparks@bristol.ac.uk
[2]De Beers MRM Group, Wells, Somerset BA5 3DG, UK.
[3]Department of Mechanical Engineering, University of Bristol, Bristol BS8 1TR, UK.

Received 26 July 2007; accepted 16 October 2007.

1. Wilson, L. & Head, J. W. III. An integrated model of kimberlite ascent and eruption. *Nature* **447,** 53–57 (2007).
2. Sparks, R. S. J. *et al.* Dynamics of kimberlite volcanism. *J. Volcanol. Geotherm. Res.* **155,** 18–48 (2006).
3. Lister, J. R. & Kerr, R. C. Fluid-mechanical models of crack propagation and their application to magma transport in dykes. *J. Geophys. Res.* **96,** 10049–10077 (1991).
4. Menand, T. & Tait, S. R. The propagation of a buoyant liquid-filled fissure from a source under constant pressure: an experimental approach. *J. Geophys. Res.* 107, 2306 16–1-14 (2002).
5. Brooker, R. A., Kohn, S., Holloway, J. R. & McMillan, P. F. Structural controls on the solubility of CO_2 in silicate melts. Part I: bulk solubility data. *Chem. Geol.* **174,** 225–239 (2001).
6. Mitchell, R. H. *Kimberlites: Mineralogy, Geochemistry and Petrology* (Plenum, New York, 1986).
7. Menand, T. & Tait, S. R. A phenomenological model for precursor volcanic eruptions. *Nature* **411,** 678–680 (2001).
8. Brown, R. J., Kavanagh, J., Sparks, R. S. J., Tait, M. & Field, M. Mechanically disrupted and chemically weakened zones in segmented kimberlite dike systems cause the localisation of kimberlites. *Geology* **35,** 815–818 (2007).
9. McLeod, P. & Tait, S. R. The growth of dykes from magma chambers. *J. Volcanol. Geotherm. Res.* **92,** 231–245 (1999).
10. Fedortchouk, Y. & Canil, D. Intensive variables in kimberlite magmas, Lac de Gras, Canada and implications for diamond survival. *J. Petrol.* **45,** 1725–1745 (2004).
11. Stripp, G., Field, M., Schumacher, J. C. & Sparks, R. S. J. Post-emplacement serpentinisation and related hydrothermal metamorphism in a kimberlite from Venetia, South Africa. *J. Metamorph. Geol.* **24,** 515–534 (2006).
12. Barnett, W. The rock mechanics of kimberlite pipe formation. *J. Volcanol. Geotherm. Res.* (in the press).
13. Walters, A. L. *et al.* The role of fluidisation in the formation of volcaniclastic kimberlite: grain size observations and experimental investigation. *J. Volcanol. Geotherm. Res.* **155,** 119–137 (2006).
14. Gernon, T., Gilbertson, M. A., Sparks, R. S. J. & Field, M. Gas-fluidisation in an experimental tapered bed: insights into processes in diverging volcanic conduits. *J. Volcanol. Geotherm. Res.* (in the press).
15. Gurney, J. in *Kimberlites and related rocks: their mantle/crust setting, diamonds and diamond exploration. Proceedings of the Fourth International Kimberlite Conference.* (eds Ross, J. *et al.*) *Geol. Soc. Australia Special Publication*, **14,** 935–965, 990–1000 (Perth, Australia, 1989).

doi:10.1038/nature06435

Wilson & Head reply

Replying to: R. S. J. Sparks, R. J. Brown, M. Field & M. Gilbertson *Nature* **450,** doi:10.1038/nature06435 (2007).

Differences between the model of Sparks *et al.*[1,2] and ours[3] arise mainly because we focus on phenomena during the transient, opening phase that we suggest dominates many kimberlite eruptions, rather than on the subsequent, more prolonged phases relevant to other kimberlite eruptions[1,2].

If water dominates carbon dioxide (CO_2) as the vapour phase[1,2], our argument[3] about the pressure distribution driving a kimberlite dyke to the surface is reinforced. The key factor allowing the initial rapid ascent is the large difference between the high mantle source pressure and the low dyke tip pressure, the latter being buffered by the saturation pressure of the least soluble volatile phase[4,5]. The dyke tip pressure required for water to exsolve will be even lower than the pressure we inferred for CO_2, thus increasing the pressure difference driving the magma upward through the opening dyke.

Any vapour-filled region at the tip of a dyke, breaking away to propagate faster as an independent crack, can incorporate[6] some of the magmatic foam implied by our model[3]. Cracks longer than ~20 m travel at ~1 km s^{-1}, which is ~40% of the sound speed in rock[7]. Chilling of magma in the closing crack base 'heals' the fracture, restoring the country rock mechanical properties. When the dyke tip subsequently arrives, it encounters essentially the same conditions as if crack separation had never occurred. Only seconds are needed to chill a 1–2-mm-thick film of magma left behind by a ~20-m-long crack; during this time the dyke tip, rising[3] at ~20 m s^{-1}, travels ~100 m—a tiny fraction of the dyke's vertical extent. A new low-pressure region starts to grow below the dyke tip immediately after crack separation; we infer that the stress and pressure conditions we proposed will be present over most of the path of the rising dyke tip.

Our dyke geometries are only slightly larger than those of Sparks *et al.*[1,2], and their minimum estimates of total magma volumes imply eruption durations only a few times longer than the time to establish the dyke pathway[3]; larger volumes will imply more prolonged events. Our calculations[3] of adiabatic cooling refer to magma reaching the surface during the opening phase of an eruption; in a long-lived eruption, most of the magma finally emplaced in the sub-surface diatreme will indeed suffer less cooling.

We suggested[3] a violent change from overpressure to underpressure as a dyke reached the surface, with rapid physical development of the near-surface pipe and diatreme system. A longer-lived eruption[1,2] will indeed allow a range of additional failure mechanisms. Although 'fluidization' commonly relates to the near-steady passage of gas through unconsolidated granular materials, as in the waning phases of kimberlite eruptions[8], the basic physics is the same as that in our violent opening phase.

Regarding preservation of diamonds during transit to the surface, we stress[3] that rapid transport will maximise the survival of diamonds as they pass through potentially unstable combinations of ambient pressure and temperature conditions, irrespective of the chemical environment that they encounter[9].

Lionel Wilson[1] & James W. Head III[2]
[1]Environmental Science Department, Lancaster University, Lancaster LA1 4YQ, UK.
[2]Geological Sciences Department, Brown University, Providence, Rhode Island 02912, USA.
e-mail: james_head@brown.edu

1. Sparks, R. S. J. *et al.* Dynamics of kimberlite volcanism. *J. Volcanol. Geotherm. Res.* **155,** 18–48 (2006).
2. Sparks, R. S. J., Brown, R. J., Field, M. & Gilbertson, M. Kimberlite ascent and eruption. *Nature* **450,** doi:10.1038/nature06435 (2007).
3. Wilson, L. & Head, J. W. III. An integrated model of kimberlite ascent and eruption. *Nature* **447,** 53–57 (2007).
4. Lister, J. R. & Kerr, R. C. Fluid-mechanical models of crack propagation and their application to magma transport in dykes. *J. Geophys. Res.* **96,** 10049–10077 (1991).
5. Rubin, A. M. Dikes vs. diapirs in viscoelastic rock. *Earth Planet. Sci. Lett.* **119,** 641–659 (1993).
6. Menand, T. & Tait, S. R. A phenomenological model for precursor volcanic eruptions. *Nature* **411,** 678–680 (2001).
7. Dobran, F. *Volcanic Processes — Mechanisms in Material Transport* page 212 (Kluwer/Plenum, New York, 2001).
8. Walters, A. L. *et al.* The role of fluidisation in the formation of volcaniclastic kimberlite: grain size observations and experimental investigation. *J. Volcanol. Geotherm. Res.* **155,** 119–137 (2006).
9. Ross, J. *et al.* (eds) *Kimberlites and related rocks: their mantle/crust setting, diamonds and diamond exploration. Proceedings of the Fourth International Kimberlite Conference.* Geol. Soc. Australia Special Publication, **14,** 935–965, 990–1000 (Perth, Australia, 1989).

doi:10.1038/nature06436

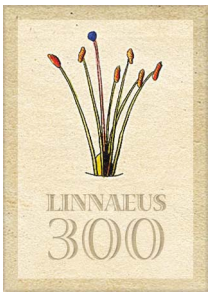
REVIEWS

Transformation and diversification in early mammal evolution

Zhe-Xi Luo[1]

Evolution of the earliest mammals shows successive episodes of diversification. Lineage-splitting in Mesozoic mammals is coupled with many independent evolutionary experiments and ecological specializations. Classic scenarios of mammalian morphological evolution tend to posit an orderly acquisition of key evolutionary innovations leading to adaptive diversification, but newly discovered fossils show that evolution of such key characters as the middle ear and the tribosphenic teeth is far more labile among Mesozoic mammals. Successive diversifications of Mesozoic mammal groups multiplied the opportunities for many dead-end lineages to iteratively evolve developmental homoplasies and convergent ecological specializations, parallel to those in modern mammal groups.

Mammals are an important group for understanding life and its evolution. With some 5,400 extant species and 4,000 fossil genera, they developed a spectacular diversity of ecomorphological specializations, ranging from the 1-gram bumblebee bat to the 100-tonne blue whale. Basal diversifications of the three extant mammalian groups, monotremes (egg-laying mammals), marsupials (pouched mammals) and placentals, occurred in the Mesozoic Era[1–4]. Their ancestors are nested in a great evolutionary bush with 25 or so lineages that co-existed with non-avian dinosaurs and other small vertebrates during the Mesozoic. Mammals were not abundant in the Mesozoic, but they were relatively diverse. Compared to the 547 known dinosaur genera[5], over 310 Mesozoic mammaliaform genera are now known to science, two-thirds of which were discovered in the last 25 years (Box 1).



The rise of mammals from premammaliaform cynodonts is an important transition in vertebrate evolution[1,2,6–9]. This already richly documented transition has been rapidly re-written by recent discoveries of very informative fossils (Box 1), by the increasingly comprehensive phylogenies with which to infer the pattern of diversification (Fig. 1), and by a more complex picture of the evolution of key anatomical features. The newly improved fossil record can reciprocally illuminate the molecular evolution of mammals, especially in light of the large discrepancies between the molecular time estimates and the fossil records for the origins of major marsupial and placental super-order lineages. These new fossils and their analyses shed new light on several controversies:

- Temporal evolution: is early mammal evolution best characterized by major long branches reaching deep into the Mesozoic and by the long evolutionary fuse that delayed diversification within long branches? Or is this evolution dominated by many short-lived branches with a short evolutionary fuse before diversification?
- Ecological diversification: is lineage splitting of early mammals decoupled from, or correlated with ecological diversification?
- Morphological transformation: are originations of key mammalian characters singular evolutionary events, or iterative convergences despite their complexity?

Temporal pattern of early mammal evolution

The evolution of early mammals occurred in successive diversifications or episodes of quick splitting of relatively short-lived clades. Most order- or family-level clades are clustered around the several nodes of their evolutionary tree. Mapped on the geological time scale, successive clusters of emergent clades represent waves of diversification (Fig. 1). Clades in a preceding episode of diversification are mostly dead-end evolutionary experiments; the majority of them have no direct ancestor–descendant relationship to the emergent clades in the succeeding episode of diversification, consistent with significant taxonomic succession and turnover between mammaliaform faunas of different geological epochs. The main episodes of diversification are: diversification of premammalian mammaliaforms—the extinct relatives outside mammals—during the Late Triassic and Early Jurassic (Fig. 1, node 1), the Middle Jurassic diversification of docodonts, theriiform mammals, and the australosphenidan mammals that are basal to monotremes (Fig. 1, node 2), the Late Jurassic diversification within the extinct theriiform groups (Fig. 1, node 3) that are closer to marsupials and placentals than to monotremes (Fig. 1, node 4), and the Early Cretaceous divergence of the marsupial lineage and the placental lineage (Fig. 1, nodes 5 and 6).

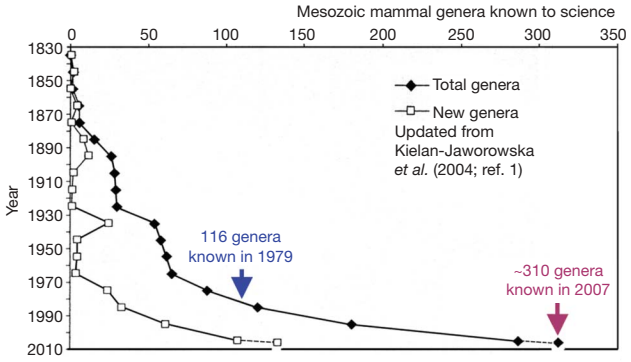
Cenozoic placentals and marsupials represent a new episode of diversification in succession to the Cretaceous stem eutherians and metatherians. Cenozoic marsupials are nested, as a whole, in the Cretaceous metatherians, but the emergent Cenozoic marsupial orders or families cannot be related directly to the known Cretaceous metatherian genera by the best available morphological data sets[10–12]. The latest analyses of all eutherians also strongly favour placement of all known eutherians of the Cretaceous outside the Cenozoic placentals[13,14], in contrast to a previous analysis[15]. The successive clusters of emergent clades and faunal turnover between the Cretaceous and Cenozoic are consistent with the overall pattern of successive diversification of Mesozoic mammaliaforms as a whole (Fig. 1).

These prevailing patterns are significantly different from the historical but now out-of-date views that a few long branches of Cenozoic or extant mammals would extend deep into the Mesozoic, but taxonomic diversification would be confined in a

[1]Carnegie Museum of Natural History, Pittsburgh, Pennsylvania 15213, USA.

Box 1 | Rapid accumulation of new data by recent discoveries of Mesozoic mammals.

In comparison to 116 Mesozoic mammal genera known to science in 1979 (ref. 3), about 200 additional Mesozoic mammals were discovered in the last 25 years, a tenfold increase from all those found in the first 200 years since the first Mesozoic mammal was unearthed in 1764 (ref. 1). Total Mesozoic mammal genera now number over 310, as compared to 540 co-existing dinosaur genera[5]. More important than the great increase in taxonomic abundance is the superb quality of new fossils that reveals a richer and more complex picture of their morphological evolution, and a much better data set for estimating phylogeny. Before 1990, skulls and skeletons were described only for a handful of Mesozoic mammals[3,35,54]. The best data set for estimating the early mammal phylogeny in 1988 scored 1,125 cells in the taxon-character matrix[7]. Today, at least 18 Mesozoic mammals are represented by nearly complete skeletons and twice as many by well-preserved skull fossils. The latest data sets for morphological phylogenetic estimates have scored 22,000 to 25,000 cells in matrices[20,21]—about 200 times that of the best available data set in the 1980s.



Evidence	Matrix size (taxa × characters)	Scored cells (matrix size – missing data)
Rowe (1998; ref. 7)	1,264	1,125
Luo & Wible (2005; ref. 20)	40,512	22,576
Luo *et al.* (2007; ref. 21)	45,938	25,849

few long-established lineages[16]. A widely accepted view, when only teeth were available for inferring early mammal history in the 1970s, was that two 'prototherian' and 'therian' lines extended to the Late Triassic[3]. These historical ideas are now replaced by more detailed phylogenies, with better sampling of skull and skeletal characters[7,17,18], in addition to dental evidence, in a great many more taxa[19–21] (Box 1). It is uncommon for any Mesozoic group to maintain a long history with little diversity, or a much delayed diversification within a lineage. Instead of a few long lineages, early mammal evolution has many short lineages in successive clusters (Fig. 1)[1,19,22].

Recent molecular dating of early mammal evolution also postulates the extension of long lineages of extant mammal superorders or orders deep into the Cretaceous, although for entirely different reasons. Molecular datings of the origins of major placental and marsupial clades at super-order or order levels are generally older than the earliest fossil records of these groups[23–25]. By one recent molecular estimate[26], all 18 extant placental orders originated in the early Late Cretaceous, as did two marsupial orders. The molecular picture of mammal evolution is a massive case of multiple long branches extending far back into deep history, with long-delayed diversification within each long branch, almost down to every modern placental order (for example, ref. 26).

The first appearance of a lineage in the fossil record represents its minimal age constraint. The actual origin of a lineage should be older than its earliest fossil record, given that the earliest history may not have been documented owing to an imperfect fossil record[27]. The inferred long delay of diversification within a major clade after its origination is aptly characterized as a 'long-fuse' evolution[28].

It is a matter of course that the minimum age constraint of the fossil record differs from the actual origin, but there is a great disagreement about the magnitude of this difference, or how frequently a long evolutionary fuse would occur in early mammal lineages. The older molecular dates would predict an abundance of long branches, and a long delay of diversification within each long branch after a branch's origin. However, studies using morphological data of both fossil and extant taxa demonstrate that there are few[15] or no such lineages with a long evolutionary lag time[13,14]. This discrepancy is so systemic and widespread that it cannot be explained by the difference between minimum age constraint (represented by actual fossils) and the timing of origin that can be hypothetically estimated by molecules in marsupial and placental evolution. The diversification models that have fully accounted for the incompleteness of the fossil record suggest that these discrepancies cannot be dismissed as a general artefact of an incomplete fossil record[29,30]. The latest morphological studies with nearly exhaustive sampling of Cretaceous fossils[10–14,20] have all shown significant gaps in the 'younger' fossil record compared to the much 'older' molecular dating of the marsupial and placental lineages, a phenomenon with which molecular evolutionists also agree.

To account for these broad discrepancies between the dating by fossils and the estimate by molecules, some have extensively argued that lineages could phylogenetically diverge long before their morphological diversification[31]. The putative long delay in evolution of identifiable features for fossils to demarcate the lineage's first appearance would be due to the decoupling of speciation and ecomorphological adaptation. More generally, it is proposed that splits of early mammal lineages were not accompanied by morphological differences and were 'silent' with regard to their ecological diversification[32].

Ecological diversification in Mesozoic mammals

Whether or to what extent the lineage splitting is correlated with morphological and ecological diversification is a question with broad implications for macroevolution[31–33]. Marsupials and placentals, the two main groups that make up 99% of all extant mammal species, show great ecomorphological diversity, and most of their orders have unique ecological specializations correlated with distinctive morphological traits (Fig. 2). There is no question that this spectacular ecomorphological diversification accelerated in an Early Cenozoic adaptive radiation of mammals into the niches vacated on the extinction of non-avian dinosaurs.

However, in the absence of contrary evidence from the previously poor fossil record, it was extrapolated to a broad generalization that Mesozoic mammals failed to develop any ecomorphological specializations. They were viewed as small animals with a generalized feeding and unspecialized limb structure for terrestrial habits (Fig. 2a), and without the widely divergent ecological specializations of Cenozoic descendants. The postulation that many mammal lineages have extended invisibly into the Mesozoic without morphological difference[27,31,33] is dependent on the extrapolation that Mesozoic mammals as a whole were generalized and lacking ecological diversification, owing to exclusion from diverse terrestrial niches by co-existing dinosaurs and other small vertebrates.

The hypothesis of the decoupling of phyletic divergence from ecological diversifications rests on the assumption that the major Mesozoic mammal groups lacked ecological specializations, other than generalized habits. This assumption is now falsified by discoveries of several new Mesozoic mammals with convergences to highly specialized extant mammals (Fig. 2). Although the majority of mammals in such Mesozoic ecosystems as the Jehol Biotas[34], and some earliest mammaliaforms[35], are certainly generalized (Fig. 2a), there is now strong evidence for ecological specializations in many other clades.

Fossorial behaviour was documented by taphonomic evidence for some premammaliaform cynodonts[36], but only recently did the fossorial skeletal specializations (such as scratch digging) become known for mammaliaform lineages. This is now shown to be

widespread in multituberculates, the most abundant group in Late Jurassic and Cretaceous mammalian faunas[37]. In docodonts, the hypertrophied burrowing limb features represent an exaptation for swimming, as in modern platypus, and invasion of freshwater habitats (Fig. 2b): *Castorocauda* (Middle Jurassic) has a broad, scaly and beaver-like tail for swimming[38]; *Haldanodon* from the Late Jurassic also shows phenotypic convergence to semi-aquatic moles[39,40].

Myrmecophagian ('ant-eating' or 'termite-eating') specialization for feeding on colonial insects, along with limbs built for scratch-digging, are the defining features of several placental groups (aardvark, pangolin and armadillo) and monotremes (echidnas). The hypertrophied digging-limb features and the unique columnar and enamel-less teeth were developed in the Late Jurassic *Fruitafossor*[20] (Fig. 2d) 150 million years (Myr) ago, 100 Myr before a similar character complex evolved convergently in armadillos and aardvarks, among placentals.

Predation and scavenging on vertebrates require a larger body mass than those of generalized insectivorous mammals (20 to

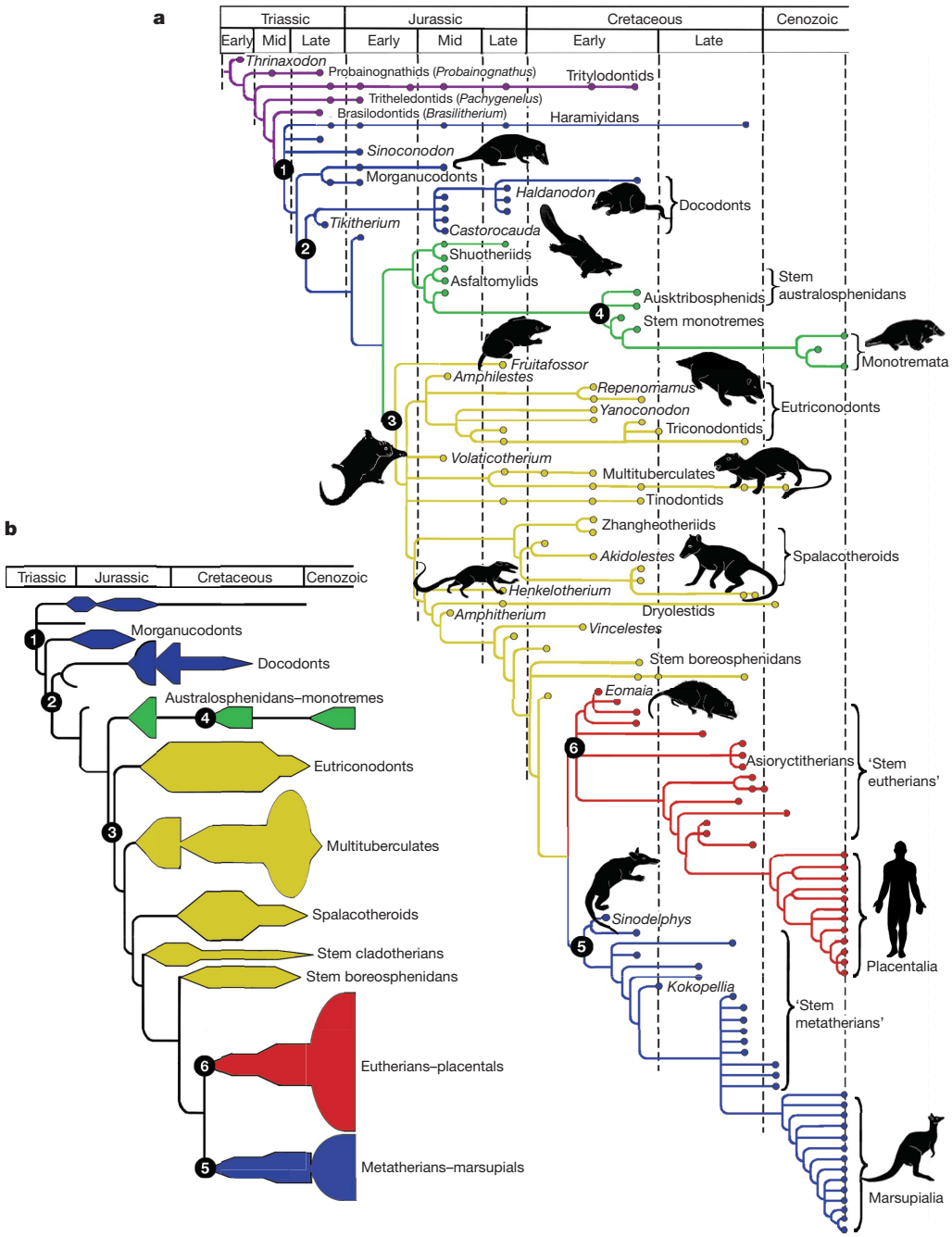


Figure 1 | Phylogeny and diversification of Mesozoic and major extant mammal groups. Almost all Mesozoic mammaliaform clades are relatively short-lived, clustered in several episodes of accelerated diversification. The short branches arising in each episode of diversification are mostly phylogenetic dead-ends without ancestor–descendant relationship to the similar dead-end branches in the episodes either before or after[1,22]. **a**, Mesozoic mammalian macroevolution is by waves of diversification of relatively short-lived clades in succession or by replacement: node 1, the Late Triassic–Early Jurassic diversification of mammaliaform stem clades (blue branches and dots); node 2, diversification of docodonts (peak diversity in the Middle Jurassic) and splits of several extinct groups in Mammalia (green and yellow); node 3, the Late Jurassic diversification within eutriconodonts, multituberculates and cladotherians; and the Early Cretaceous originations of character-based monotremes (node 4), stem-based metatherians (including marsupials; node 5) and stem-based eutherians (including placentals; node 6). Animal silhouettes are major taxa, either newly discovered or re-interpreted with better fossils after the 1990s, showing previously unsuspected ecological diversification. **b**, Diversity patterns of the order- or family-level Mesozoic mammal groups. Phylogeny is from refs 20 and 21, with additional taxa[49,57].

100 g). Some individuals could reach 500 g in *Sinoconodon*[41], 700 g in *Castorocauda*[38] and even 5–12 kg for several gobiconodontid species that could prey on other small vertebrates[42]. The Jurassic and Cretaceous saw multiple evolutions of predatory carnivores in unrelated groups (Fig. 2c).

The capacity to climb on uneven terrain is inherent in generalized small mammals[35]. Derived scansorial adaptation is widespread among Early Cenozoic marsupials[43–45] and some multituberculates[46]. New skeletal fossils suggest that some (although not all) Mesozoic eutherians and metatherians and their near kin also developed such adaptations, as shown by the elongate intermediate phalanges and convex profiles of manual and pedal distal phalanges, and in the tarsus, among other skeletal features[11,47,48] (Fig. 2e). The adaptation for climbing is a pre-requisite for extant volant (gliding and flying) mammals. The recent discovery of *Volaticotherium* (possibly a eutriconodont) shows the skin membrane (patagium) associated with elongate limbs for gliding, convergent to marsupial sugar gliders, and 'flying' squirrels and dermopterans among placentals[49] (Fig. 2f).

Treated individually, these curious cases of convergent adaptations in extinct Mesozoic mammals represent many separate evolutionary experiments[20,37–39,49]. But taken together (Fig. 2), they unveiled a new picture in which ecological diversification is not unique to the Early Cenozoic mammalian radiation, and that many dead-end Mesozoic mammal clades developed similar ecomorphotypes long before the analogous modern mammals (Fig. 2).

Although far less abundant numerically in the Mesozoic than in the Cenozoic, within the limited snap-shot windows of the Middle Jurassic to the Early Cretaceous—for which we happen to have sufficient fossil data—mammalian ecological specializations attained nearly the same diversification as the early-middle Palaeocene placentals in North America (except for cursorial ungulates) and as marsupials of the Oligocene-Miocene of Australia. The decoupling hypothesis can certainly be rejected as a rationale for the gap between molecular time estimates and the first appearance in fossil data of the major placental and marsupial lineages. Correlation of ecomorphological specializations with phylogenetic splitting is a basic feature of Mesozoic mammal evolution. Cenozoic placental carnivores are an independent case for correlated ecomorphological and phyletic diversifications[33].

Transformation of key evolutionary apomorphies

On the broadest possible scale, evolution from premammalian synapsids to mammaliaforms shows incremental acquisition of mammalian apomorphies[8,9,41,50]. Stepwise assembly of incremental precursor conditions towards complex mammal structure is an evolutionary paradigm of functional adaptation and taxonomic diversification of mammals[50–52]. Some best-documented 'textbook' scenarios are acquisitions of key characters along a transformation series: transformations of the mammalian middle ear and the jaw hinge (Fig. 3), and evolution of the tribosphenic molars (Fig. 4).

Homoplasies in mammal middle-ear evolution. The postdentary bones in the posterior part of the mandible make up the jaw hinge and the mandibular middle ear in premammalian cynodonts. They show a gradual size reduction in the mandible—as the dentary bone shows gradual enlargement—among transitional taxa successively

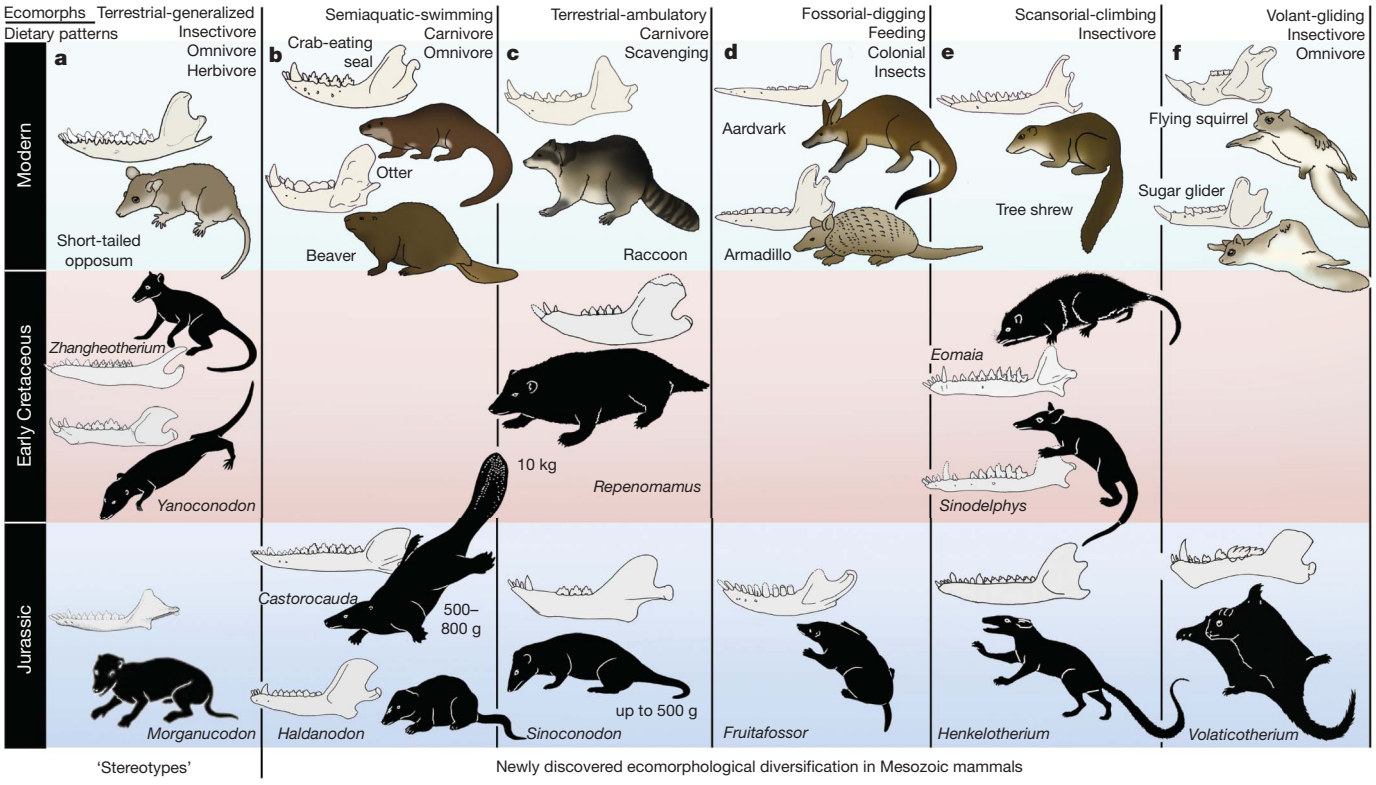


Figure 2 | Diverse evolutionary experiments of Mesozoic mammals and their ecological convergence to modern mammal ecomorphotypes. **a**, Representation of the traditional assumption that Mesozoic mammals were generalized small animals with generalized feeding and terrestrial habits, and had few of the diverse ecomorphotypes of Cenozoic mammals; the hypothesis on decoupling of lineage splitting from ecological diversification is based on this assumption[27,31,32], which is now contradicted by recent discoveries of a great range of ecological specializations, such as: **b**, swimming and fish-feeding in the docodont *Castorocauda*[38], and semi-aquatic habits of *Haldanodon*[39,40]; **c**, ambulatory carnivory or scavenging (predation or feeding on other vertebrates) in large gobiconodontids[42] and large individuals of *Sinoconodon*[41]; **d**, scratch-digging and feeding on colonial insects in *Fruitafossor*[20]; **e**, scansorial (climbing) limb characteristics in basal eutherians and metatherians, and their near relatives[11,47,48]; and **f**, volant (gliding) adaptation in *Volaticotherium*[49]. The Jurassic and Cretaceous mammals developed, iteratively, similar niche specializations to modern Australasian monotremes and marsupials, and are no less diverse, ecologically, than the early-to-middle Palaeocene mammals of similar body-size range. Splits of Mesozoic mammal groups were accompanied by ecological diversification.

closer to mammals (Fig. 3a–e)[8,52–56]. In more derived premammalian mammaliaforms, the dentary is so enlarged as to have a condyle articulating with the squamosal glenoid, forming the true mammalian jaw hinge, known as the temporomandibular joint (Fig. 3d, e). Further along the evolution of living mammals, the middle ear became detached from the mandible to form the 'cranial middle ear', or the definitive mammalian middle ear (Fig. 3f, h). The detachment of the middle ear from the mandible in adults and the mobile suspension of the middle ear via the incus to the cranium are crucial for sensitivity of the mammal middle ear[55,56]. Sound transition from the tympanic membrane through the middle ear also requires the malleus manubrium, as an in-lever, and the incus stapedial process, as an out-lever, for the impedance-match and amplification of airborne sound (Fig. 3)[51,55].

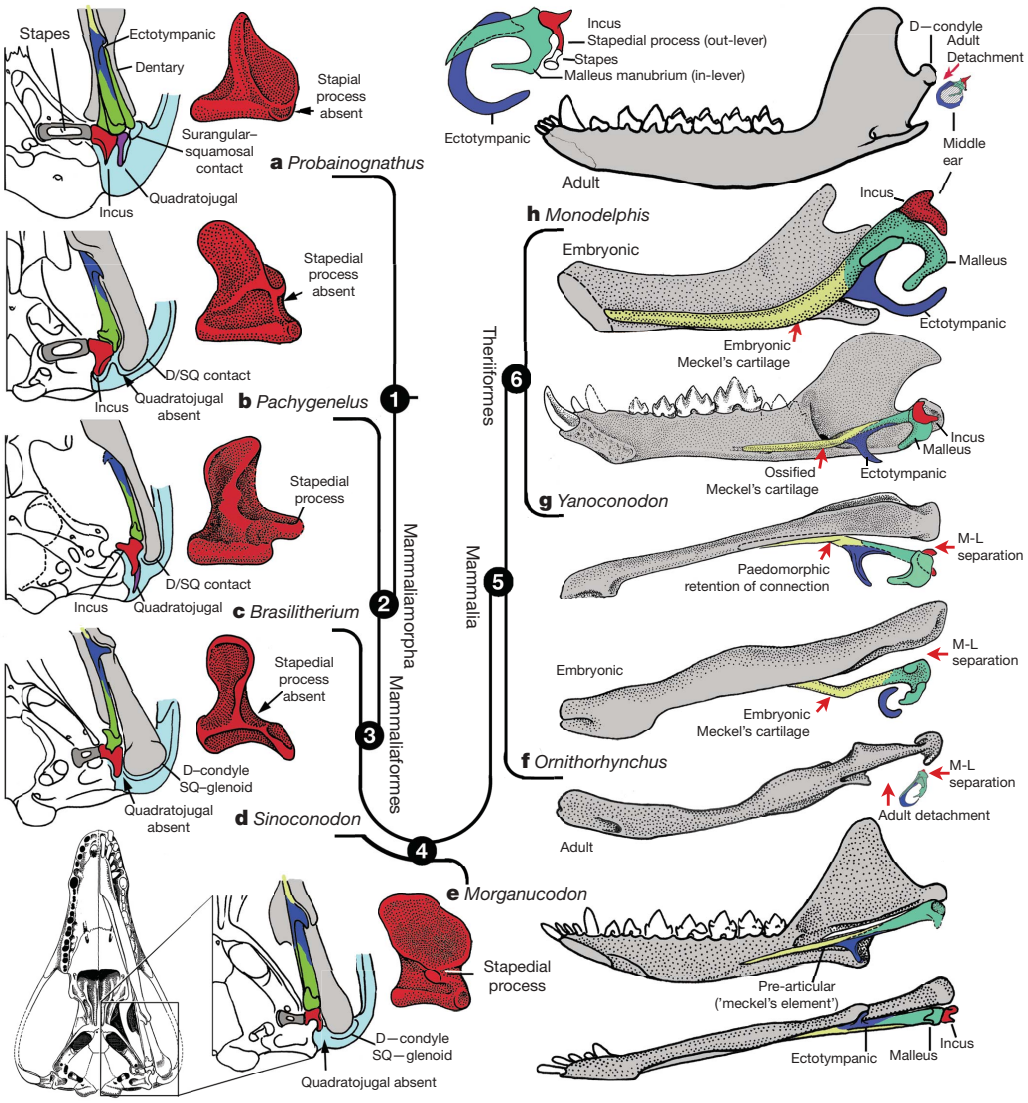


Figure 3 | Evolution of the mammalian cranio-mandibular joint and the definitive mammalian middle ear through the cynodont–mammal transition. Homoplasies occurred for the simplification of the incus articulation, the stapedial process of the incus and the detachment of the ectotympanic from mandible. **a**, The cynodont *Probainognathus*[51,55]; ventral view of left basicranium and posterior view of the incus (quadrate). **b**, The mammaliamorph *Pachygenelus*[51,55]. **c**, The mammaliamorph *Brasilitherium* (modified from ref. 57 by personal observation). **d**, The mammaliaform *Sinoconodon*. **e**, *Morganucodon* (redrawn from refs 53 and 54): left panel, left basicranium, ventral view; middle panel, left incus, posteromedial view; and right panel, the mandible and 'mandibular' middle ear in ventral (below) and medial (above) views. **a**–**e**, Homoplastic loss of the quadratojugal for a more mobile incus occurs in *Pachygenelus* (**b**) and mammaliaforms (**d**, **e**), but not in *Probainognathus* (**a**), tritylodontids (not shown) and *Brasilitherium* (**c**). The stapedial process of the incus, the out-lever for the middle ear, is present in tritylodontids (not shown), *Brasilitherium* (**c**) and *Morganucodon* (**e**), but not in other taxa (**a**, **b**, **d**). **f**, The monotreme *Ornithorhynchus* lower jaw (ventral view): the middle ear attached anteriorly to the mandible by Meckel's cartilage in the embryonic stage[59], but detached from the mandible after re-absorption of Meckel's cartilage in the adult. **g**, The eutriconodont *Yanoconodon* lower jaw (lower panel, ventral view; upper panel, medial view): the middle ear is medio-laterally (M-L) separated from, but anteriorly connected to, the mandible by the prematurely ossified Meckel's cartilage, similar to the embryonic condition of monotremes of medio-lateral (M-L) separation of the ear from the mandible, and to the monotreme configuration of the ectotympanic and malleus. **h**, The medial view of the mandible and middle ear of the marsupial *Monodelphis*: the middle ear is attached to the mandible by Meckel's cartilage in the embryonic stage[60,61], but detached from the mandible after the re-absorption of Meckel's cartilage in the adult. Because *Yanoconodon* (**g**) is nested between extant monotremes (**f**) and therians (**h**), both of which have have separation of the middle ear from the mandible, the Meckel's connection of the ectotympanic to the mandible in *Yanoconodon* shows that some Mesozoic mammals had homoplastic evolution of the definitive mammalian middle ear, defined by full detachment of the ectotympanic from the dentary. The ossified Meckel's cartilage of *Yanoconodon* is very similar to the embryonic Meckel's cartilage of extant monotremes, and has paedomorphic resemblance to the embryonic condition of extant mammals. The homoplastic attachment of the mandible and the middle ear in *Yanoconodon* is correlated with changes in the timing and rate of development. D, dentary; SQ, squamosal; D/SQ, the dentary–squamosal contact or joint.

If mapped on a limited number of exemplary fossils on a broad phylogenetic scale, evolution of the definitive mammalian middle ear and mammalian jaw hinge is orderly both in qualitative[51,55] and quantitative terms[8,56]. However, a series of newly discovered fossils have shown more complex transformations of the main components of the mammalian middle ear[21,57]. This can be demonstrated for how the middle ear became connected to the cranium but disconnected from the mandible.

Mobile suspension of the middle ear and its impedance-match system. A highly agile and mobile suspension of the incus in the cranium contributes to sensitive hearing function. The incus is ancestrally associated with the quadratojugal bone (Fig. 3a, c, purple). The quadratojugal–incus articulation to the cranium reinforces the incus for the load-bearing function of the jaw hinge, but also reduces the hearing sensitivity (Fig. 3a, c). The stapedial process is present in most mammaliaforms, fulfilling the crucial function of the out-lever of the middle-ear lever system for the impedance-match and amplification, but is absent in the incus of most premammaliaform cynodonts (Fig. 3a, b). These functionally important apomorphies (and their respective precursory conditions) have incongruent distributions in the transitional taxa from premammalian cynodonts to mammaliaforms. The stapedial process has a discontinuous distribution: it is present in tritylodontids (not illustrated), *Brasilitherium*[57] and *Morganucodon*[17,51,54], but absent in the tritheledontid *Pachygenelus*[51]—which is phylogenetically between tritylodontids and *Brasilitherium*—and in *Sinoconodon*, which is between *Brasilitherium* and *Morganucodon*. The quadratojugal is lost in *Pachygenelus*, *Sinoconodon* and *Morganucodon*, thereby allowing more mobility in the middle ear, but is present in *Brasilitherium*, a taxon more derived than *Pachygenelus* in cynodont–mammal evolution[57].

Regardless of the alternative tree topology of such transitional forms as *Pachygenelus*, *Brasilitherium* and *Sinoconodon*, the loss of the quadratojugal and the development of the stapedial process are not only homoplastic in their overall distributions, but are also in conflict with each other. It is abundantly evident that separate evolutionary experiments occurred repetitively during the transformation of the incus structure for better impedance-match and hearing sensitivity.

Mandible–ear detachment and formation of the definitive mammalian middle ear. In premammalian outgroups, the middle ear is attached to the dentary, by the pre-articular (an ossified Meckel's element) and the ectotympanic (Fig. 3e). In extant mammals, such as the monotreme *Ornithorhynchus* (Fig. 3f) and the marsupial *Monodelphis* (Fig. 3h), the connection between the dentary and Meckel's element is conserved in embryonic and fetal stages, but lost in the adult owing to the re-absorption of embryonic Meckel's cartilage, the homologue to part of the prearticular[56,58–62].

Opinions waxed and waned as to whether detachment of the definitive mammalian middle ear occurred a single time[7,17,56,63,64] or more than once in mammal evolution[21,53,65,66]. It can be argued that disconnection by the adult re-absorption of the embryonic Meckel's cartilage happened only once, and that the definitive mammalian middle ear had a monophyletic origin, if these extant mammals are directly compared to the premammalian mammaliaforms without considering several fossil groups nested within the crown Mammalia. Adult monotremes have complete separation of the middle ear from the mandible (Fig. 3f), but in extinct taxa in the monotreme

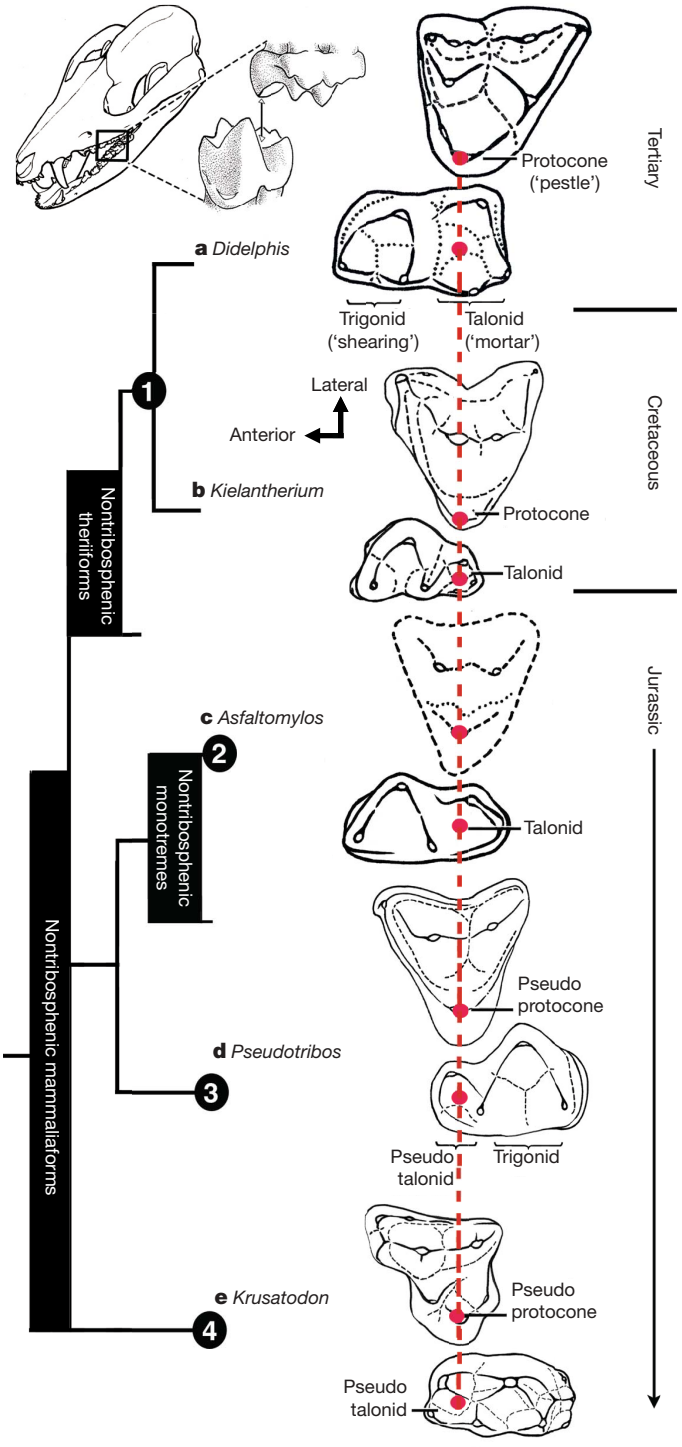


Figure 4 | Convergent and iterative evolution of protocones and pseudoprotocones in Mesozoic mammals. The tribosphenic and pseudotribosphenic molars achieved analogous pulping, crushing and grinding functions by opposite arrangements of main structures: in tribosphenic molars the protocone of the upper molar is aligned to the talonid basin posterior to the primitive trigonid of the lower; in pseudotribosphenic molars the analogous pseudoprotocone is aligned to a pseudotalonid basin anterior to the same trigonid on the lower molar. The protocone or its analogous cusp is developed independently from the immediate ancestors without such a structure (black bands) four times: boreosphenidan mammals (node 1, *Kielantherium* + the common ancestor of marsupials and placentals), australosphenidans (node 2, Gondwanan tribosphenic mammals as the immediate outgroups to non-tribosphenic monotremes), pseudotribosphenidans (node 3, *Pseudotribos* and kin) and docodont mammals (node 4, *Krusatodon*). Three lineages had experimented with the protocone or a similar structure in the Middle Jurassic without success, and gone extinct, long before the common ancestor of marsupials and placentals re-evolved the protocone, which may be correlated to their early diversification. **a**, The marsupial *Didelphis* had typical tribosphenic molars. **b**, The Early Cretaceous northern tribosphenic (boreosphenid) *Kielantherium*[74]. **c**, The Middle Jurassic southern tribosphenic (australosphenid) *Asfaltomylos* (hypothetical upper molar)[82,83]. **d**, The Middle Jurassic pseudo-tribosphenic (shuotheriid) *Pseudotribos*[78]. **e**, The Middle Jurassic 'pseudo-tribosphenic' docodont *Krusatodon*[90].

lineage, the receiving structure on the mandible for connecting the middle ear is still present[19,65], although the middle-ear bones themselves are not preserved, causing some uncertainties in interpretation[63,64]. For groups that are nested among modern mammals, the most conclusive evidence for attachment of the middle ear to the mandible is from several eutriconodonts. Several gobiconodontids have preserved an ossified Meckel's cartilage[67–69]. In the newly discovered *Yanoconodon*[21], this ossified Meckel's cartilage connects the mandible to the ectotympanic and the malleus, the two bones supporting the tympanic membrane in extant mammals. Regardless of whether the middle ear's connection to the mandible is considered to be an atavistic reversal or a convergent acquisition, it is beyond doubt that the last step in the transformation of the definitive mammalian middle ear occurred homoplastically in some Mesozoic lineages (Fig. 3).

The ossified Meckel's cartilage of eutriconodonts is morphologically similar to the embryonic Meckel's cartilage of extant monotremes in having a bend in the Meckel's cartilage and in the medio-lateral (M-L) separation of the ectotympanic and malleus from the mandible, and can be regarded to be paedomorphic by comparison to the embryonic condition of extant monotreme and placental mammals[59–61] (Fig. 3). The middle ear's attachment to the mandible in *Yanoconodon* (and possibly in eutriconodonts as a whole) is attributable to differences in developmental timing and rate between *Yanoconodon* and extant mammals. Because reabsorption of Meckel's cartilage is crucial for extant mammals to complete the ontogeny of their middle ear, an early ossification of Meckel's cartilage influenced the retention of the ectotympanic–dentary connection in some major Mesozoic mammal groups. This provides a common ontogenetic heterochrony as a main mechanism for the homoplastic evolution of a critical component of the mammalian middle ear.

Evolution of tribosphenic and pseudotribosphenic molars. Tribosphenic molars of basal marsupials and placentals have the protocone (pestle) of the upper molar crushing and grinding in the talonid basin (mortar) on the lower molar[70–74]. Because this new function by the derived protocone and talonid is added to the basic shearing function of the primitive structure of the trigonid (Fig. 4), this complex structure with more versatile functions is considered to be a key dental innovation for more effective faunivory and omnivory, leading to the basal diversification of marsupials and placentals. It was widely assumed that the upper-molar protocone, the lower-molar talonid, and their occlusal correspondence evolved together in a single origin in the group Tribosphenida, defined by the common ancestor of marsupials, placentals and their proximal kin[72,73].

However, the discovery of the pseudotribosphenic mammals *Shuotherium* and *Pseudotribos* changed the assumption that the derived function of the protocone- and the talonid-like structure was a singular evolutionary event[75–78]. Pseudotribosphenic molars have a design that is geometrically reversed from that of the tribosphenic molars: a pseudo-talonid is anterior to the trigonid, and receives the pseudo-protocone of the upper molar (Fig. 4d). This functionally analogous pseudo-talonid is anteriorly placed in pseudotribosphenic mammals and opposite to the posterior talonid basin of the true tribosphenic mammals (Fig. 4a, b). Therefore, a protocone-like structure of the upper molar can occlude either a talonid in the posterior part of the lower molar, or a pseudo-talonid in the anterior part of the lower molar, in different clades; the protocone-like structure of the upper molar evolved homoplastically in mammalian history.

Discoveries of southern tribosphenic mammals, or australosphenidans, from the Mesozoic of Gondwana[79–84] falsified the traditional notion that tribosphenic mammals had a single origin on the northern continents[72,73]. The earliest tribosphenic mammals of Gondwana are fairly diverse, with a wide distribution. They are more derived than the northern tribosphenic mammals with respect to unique premolar features and in having distinctive wear patterns concentrated apically on the peripheral crests of the molar talonid; this is similar to toothed monotremes, but not boreosphenidans[83–85]. One school of thought argues that these australosphenidans are placentals[79,80,87]. Because australosphenidans have the postdentary trough accommodating the mandibular middle ear[83,84], this implies that the ancestral mandibular ear would have re-evolved independently within placentals after the marsupial–placental split. This hypothesis also postulates that placentals would originate earlier than 170 Myr ago[87], much earlier than even the current earliest molecular dating (~147 Myr ago) for the placental–marsupial split[26]. A contrasting view, based on analyses of all major Mesozoic and extant mammal clades (Fig. 1), is that the lower-molar talonid basin in australosphenidans represents convergent evolution. These southern mammals are extinct relatives to monotremes, which are relictual taxa from an ancient mammal diversification within the Gondwanan continents[1,19,85]. Several recent and independent analyses supported the hypothesis of dual evolution of tribosphenic molars and the australosphenidan clade[83,84,88], some with modified outgroup relationships of australosphenidans[84,89].

Some docodont mammaliaforms also achieved a pseudo-protocone structure not unlike those of tribosphenic or pseudotribosphenic molars[90]. Three Middle Jurassic lineages developed a protocone, or a similar structure, without much evolutionary success, and became extinct long before the common ancestor of marsupials and placentals re-evolved the protocone during their Cretaceous and early Cenozoic diversification. Dental evolution was far more labile in Mesozoic mammals than can be inferred from Cenozoic mammals (Fig. 4)[19,83–85], and is consistent with the functional analysis that there was more than one pathway to combine slicing and crushing functions, as exemplified by tribosphenic and pseudotribosphenic molars for more effective faunivory and omnivory, in early mammalian history[91].

Concluding remarks

The traditional paradigm of early mammal evolution portrayed the origin of key innovations as an incremental assembly of complex features with great functional adaptation in the time of diversification of a major group. Two classic examples of this paradigm are the sensitive hearing by the sophisticated middle ear in the earliest mammaliaforms, leading to exploitation of the nocturnal niches, and the versatile functions of the tribosphenic molar in northern tribosphenic mammals, leading to the great diversification of marsupials and placentals. Because there used to be no evidence to the contrary, it was granted that processes of evolutionary innovation leading to ecological diversification were singular events—these evolutionary innovations of mammals are so intricate and unique that it would be unlikely for these sophisticated structures to be homoplastic[33,66].

Character conflicts are inevitable when more characters become available from better 'transitional' fossils. For the several key mammaliaform structures known to have evolved by incremental or stepwise assembly, their precursory conditions have shown character conflicts in the recently found fossils (Fig. 3). This suggests labile evolutionary experiments before the accomplishment of the complex structure. Character transformation and the attendant homoplasies can now be attributable to functional adaptation, evolutionary development, or both. Homoplasies in the definitive mammalian middle ear by the ossified Meckel's cartilage in eutriconodonts are a case of developmental heterochrony. Models on developmental mechanism[92,93] and functional analysis[91] of dental characters are consistent with iterative evolution of the protocone-like structure among docodont, pseudotribosphenic and tribosphenic mammals, as postulated by parsimonious phylogeny of fossils (Fig. 4). Other similar examples include thoraco-lumbar vertebral homoplasies among Mesozoic mammals that are dead-ringers for loss and gain of *hox* gene patterning[21,94–98]. Perhaps most interestingly, successive waves of Mesozoic mammal diversification multiplied the chances for many short-lived lineages to iteratively experiment with developmental patterning and ecological diversification that were previously

known only for Cenozoic mammals, but that are now shown to be widespread among Mesozoic mammals. This shows that lineage splits are accompanied by significant ecological diversification and by more labile developmental patterning in early mammal evolution.

An emergent new paradigm is that successive diversifications of Mesozoic mammals made it possible for many extinct lineages to exploit diverse niches—as during Cenozoic mammalian diversification (albeit less successfully)—in independent evolutionary experiments facilitated by extensive developmental homoplasies and convergent functional and ecological adaptation.

1. Kielan-Jaworowska, Z. *et al. Mammals from the Age of Dinosaurs—Origins, Evolution, and Structure* (Columbia Univ. Press, New York, 2004).
2. Kemp, T. S. *The Origin And Evolution of Mammals* (Oxford Univ. Press, Oxford, 2005).
3. Lillegraven, J. A., Kielan-Jaworowska, Z. & Clemens, W. A. (eds) *Mesozoic Mammals: The First Two-thirds of Mammalian History* (Univ. Calif. Press, Berkeley, 1979).
4. McKenna, M. C. & Bell, S. K. *Classification of Mammals Above the Species Level* (Columbia Univ. Press, New York, 1997).
5. Wang, C. S. & Dodson, P. Estimating the diversity of dinosaurs. *Proc. Natl Acad. Sci. USA* **103,** 13601–13605 (2006).
6. Hopson, J. A. & Kitching, J. W. A probainognathian cynodont from South Africa and the phylogeny of nonmammalian cynodonts. *Bull. Mus. Comp. Zool. (Harvard)* **156,** 5–35 (2001).
7. Rowe, T. B. Definition, diagnosis, and origin of Mammalia. *J. Vertebr. Paleontol.* **8,** 241–264 (1988).
8. Sidor, C. A. Simplification as a trend in synapsid cranial evolution. *Evolution Int. J. Org. Evolution* **55,** 1419–1442 (2001).
9. Sidor, C. A. & Hopson, J. A. Ghost lineages and "mammalness": assessing the temporal pattern of character acquisition in the Synapsida. *Paleobiology* **24,** 254–273 (1998).
10. Rougier, G. W. *et al.* Implications of *Deltatheridium* specimens for early marsupial history. *Nature* **396,** 459–463 (1998).
11. Luo, Z.-X. *et al.* An Early Cretaceous tribosphenic mammal and metatherian evolution. *Science* **302,** 1934–1940 (2003).
12. Asher, R. J. *et al.* First combined cladistic analysis of marsupial mammal interrelationships. *Mol. Phylogenet. Evol.* **33,** 240–250 (2004).
13. Asher, R. J. *et al.* Stem Lagomorpha and the antiquity of Glires. *Science* **307,** 1091–1094 (2005).
14. Wible, J. R. *et al.* Cretaceous eutherians and Laurasian origin for placental mammals near the K-T boundary. *Nature* **442,** 1003–1006 (2007).
15. Archibald, J. D. *et al.* Late Cretaceous relatives of rabbits, rodents, and other extant eutherian mammals. *Nature* **414,** 62–65 (2001).
16. Simpson, G. G. *A Catalogue of the Mesozoic Mammalia in the Geological Department of the British Museum* (British Museum, London, 1928).
17. Kemp, T. S. The relationships of mammals. *Zool. J. Linn. Soc.* **77,** 353–384 (1983).
18. Wible, J. R. & Hopson, J. A. in *Mammal Phylogeny* Vol. 1 (eds F. S. Szalay *et al.*) 45–62 (Springer-Verlag, New York, 1993).
19. Luo, Z.-X. *et al.* In quest for a phylogeny of Mesozoic mammals. *Acta Palaeontol. Polonica* **47,** 1–78 (2002).
20. Luo, Z.-X. & Wible, J. R. A new Late Jurassic digging mammal and early mammalian diversification. *Science* **308,** 103–107 (2005).
21. Luo, Z.-X. *et al.* A new eutriconodont mammal and evolutionary development of early mammals. *Nature* **446,** 288–293 (2007).
22. Cifelli, R. L. Early mammalian radiations. *J. Paleontol.* **75,** 1214–1226 (2001).
23. Murphy, W. J. *et al.* Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294,** 2348–2351 (2001).
24. Springer, M. S. *et al.* Placental mammal diversification and the Cretaceous-Tertiary boundary. *Proc. Natl Acad. Sci. USA* **100,** 1056–1061 (2003).
25. Nilsson, M. A. *et al.* Marsupial relationships and a timeline for marsupial radiation in South Gondwana. *Gene* **340,** 189–196 (2004).
26. Bininda-Emonds, O. R. P. *et al.* The delayed rise of present-day mammals. *Nature* **446,** 507–512 (2007).
27. Benton, M. J. & Donoghue, P. C. J. Paleontological evidence to date the tree of life. *Mol. Biol. Evol.* **24,** 26–53 (2007).
28. Archibald, J. D. & Deutschman, D. H. Quantitative analysis of the timing of the origin and diversification of extant placental orders. *J. Mamm. Evol.* **8,** 107–124 (2001).
29. Foote, M. *et al.* Evolutionary and preservational constraints on origins of biologic groups: divergence times of eutherian mammals. *Science* **283,** 1310–1314 (1999).
30. Hunter, J. P. & Janis, C. M. Spiny Norman in the Garden of Eden? Dispersal and early biogeography of Placentalia. *J. Mamm. Evol.* **13,** 89–123 (2006).
31. Easteal, S. Molecular evidence for the early divergence of placental mammals. *BioEssays* **21,** 1052–1058 (1999).
32. Bromham, L. *et al.* Growing up with dinosaurs: molecular dates and the mammalian radiation. *Trends Ecol. Evol.* **14,** 113–118 (1999).
33. Wesley-Hunt, G. D. The morphological diversification of carnivores in North America. *Paleobiology* **31,** 35–55 (2005).
34. Zhou, Z.-H. *et al.* An exceptionally preserved Lower Cretaceous ecosystem. *Nature* **421,** 807–814 (2003).
35. Jenkins, F. A. Jr & Parrington, F. R. The postcranial skeletons of the Triassic mammals *Eozostrodon, Megazostrodon* and *Erythrotherium*. *Phil. Trans. R. Soc. Lond. B* **273,** 387–431 (1976).
36. Damiani, R. *et al.* Earliest evidence of cynodont burrowing. *Proc. R. Soc. Lond. B* **270,** 1747–1751 (2003).
37. Kielan-Jaworowska, Z. & Gambaryan, P. P. Postcranial anatomy and habits of Asian multituberculate mammals. *Fossils Strata* **36,** 1–92 (1994).
38. Ji, Q. *et al.* A swimming mammaliaform from the Middle Jurassic and ecomorphological diversification of early mammals. *Science* **311,** 1123–1127 (2006).
39. Martin, T. Postcranial anatomy of *Haldanodon exspectatus* (Mammalia, Docodonta) from the Late Jurasssic (Kimmeridgian) of Portugal and its bearing for mammalian evolution. *Zool. J. Linn. Soc.* **145,** 219–248 (2005).
40. Martin, T. Paleontology: early mammalian evolutionary experiments. *Science* **311,** 1109–1110 (2006).
41. Luo, Z.-X. *et al.* A new mammaliaform from the Early Jurassic of China and evolution of mammalian characteristics. *Science* **292,** 1535–1540 (2001).
42. Hu, Y.-M. *et al.* Large Mesozoic mammals fed on young dinosaurs. *Nature* **433,** 149–153 (2005).
43. Szalay, F. S. & Sargis, E. J. Model-based analysis of postcranial osteology of marsupials from the Palaeocene of Itaboraí (Brazil) and the phylogenetics and biogeography of Metatheria. *Geodiversitas* **23,** 139–302 (2001).
44. Muizon, C. de. *Mayulestes ferox*, a borhyaenoid (Metatheria, Mammalia) from the early Palaeocene of Bolivia. Phylogenetic and palaeobiologic implications. *Geodiversitas* **20,** 19–142 (1998).
45. Argot, C. Functional–adaptive anatomy of the forelimb in the Didelphidae, and the paleobiology of the Paleocene marsupials *Mayulestes ferox* and *Pucadelphys andinus*. *J. Morphol.* **247,** 51–79 (2001).
46. Krause, D. W. & Jenkins, F. A. Jr. The postcranial skeleton of North American multituberculates. *Bull. Mus. Comp. Zool. Harv.* **150,** 199–246 (1983).
47. Krebs, B. Das Skelett von *Henkelotherium guimarotae* gen. et sp. nov. (Eupantotheria, Mammalia) aus dem Oberen Jura von Portugal. *Berliner Geowissensch. Abh.* **A133,** 1–110 (1991).
48. Ji, Q. *et al.* The earliest-known eutherian mammal. *Nature* **416,** 816–822 (2002).
49. Meng, J. *et al.* A Mesozoic gliding mammal from northeastern China. *Nature* **444,** 889–893 (2006).
50. Luo, Z.-X. in *In the Shadow of the Dinosaurs—Early Mesozoic Tetrapods* (eds N. C. Fraser & H.-D. Sues) 98–128 (Cambridge Univ. Press, Cambridge, 1994).
51. Luo, Z.-X. & Crompton, A. W. Transformation of the quadrate (incus) through the transition from non-mammalian cynodonts to mammals. *J. Vertebr. Paleontol.* **14,** 341–374 (1994).
52. Crompton, A. W. in *Studies in Vertebrate Evolution* (eds K. A. Joysey & T. S. Kemp) 231–253 (Oliver & Boyd, Edinburgh, 1972).
53. Kermack, K. A. *et al.* The lower jaw of *Morganucodon*. *Zool. J. Linn. Soc.* **53,** 87–175 (1973).
54. Kermack, K. A. *et al.* The skull of *Morganucodon*. *Zool. J. Linn. Soc.* **71,** 1–158 (1981).
55. Allin, E. F. & Hopson, J. A. in *The Evolutionary Biology of Hearing* (eds Webster, D. B. *et al.*) 587–614 (Springer, New York, 1992).
56. Rowe, T. B. Coevolution of the mammalian middle ear and neocortex. *Science* **273,** 651–654 (1996).
57. Bonaparte, J. F. *et al.* New information on *Brasilodon* and *Brasilitherium* (Cynodontia, Probainognathia) from the Late Triassic, southern Brazil. *Revist. Brasil. Paleontol.* **8,** 25–56 (2005).
58. Gaupp, E. Die Reichertsche Theorie (Hammer-, Amboss- und Kieferfrage). *Archiv. Anatomie Entwick.* **1912,** 1–426 (1913).
59. Zeller, U. Die Entwicklung und Morphologie des Schädels von *Ornithorhynchus anatinus* (Mammalia: Prototheria: Monotremata). *Abh. Senckenberg. Naturforsch. Ges.* **545,** 1–188 (1989).
60. Maier, W. Phylogeny and ontogeny of mammalian middle ear structures. *Nether. J. Zool.* **40,** 55–75 (1990).
61. Maier, W. in *Mammal Phylogeny* Vol. 1 (eds Szalay, F. S. *et al.*) 165–181 (Springer, New York, 1993).
62. Sánchez-Villagra, M. R. *et al.* Ontogenetic and phylogenetic transformations of the ear ossicles in marsupial mammals. *J. Morphol.* **251,** 219–238 (2002).
63. Bever, G. *et al.* Comment on "Independent origins of middle ear bones in monotremes and therians.". *Science* **309,** 1492a (2005).
64. Rougier G. W., Forasiepi, A. M. & Martinelli, A. G. Comment on "Independent origins of middle ear bones in monotremes and therians.". *Science* **309,** 1492b (2005).
65. Rich, T. H. *et al.* Independent origins of middle ear bones in monotremes and therians. *Science* **307,** 910–914 (2005).
66. Martin, T. & Luo, Z.-X. Paleontology: homoplasy in the mammalian ear. *Science* **307,** 861–862 (2005).
67. Wang, Y.-Q. *et al.* An ossified Meckel's cartilage in two Cretaceous mammals and origin of the mammalian middle ear. *Science* **294,** 357–361 (2001).
68. Li, C.-K. *et al.* A new species of *Gobiconodon* (Triconodonta, Mammalia) and its implication for the age of Jehol Biota. *Chin. Sci. Bull.* [English] **48,** 1129–1134 (2003).
69. Meng, J. *et al.* The ossified Meckel's cartilage and internal groove in Mesozoic mammaliaforms: implications to origin of the definitive mammalian middle ear. *Zool. J. Linn. Soc.* **138,** 431–448 (2003).

70. Patterson, B. Early Cretaceous mammals and the evolution of mammalian molar teeth. *Fieldiana. Geology* **13,** 1–105 (1956).
71. Crompton, A. W. in *Early Mammals* (eds Kermack, D. M. & Kermack, K. A.) 65–87 (Zool. J. Linn. Soc., London, 1971).
72. McKenna, M. C. in *Phylogeny of the Primates* (eds Luckett, W. P. & Szalay, F. S.) 21–46 (Plenum Publ. Corp., New York, 1975).
73. Prothero, D. R. New Jurassic mammals from Como Bluff, Wyoming, and the interrelationships of non-tribosphenic Theria. *Bull. Am. Mus. Nat. Hist.* **167,** 277–326 (1981).
74. Lopatin, A. V. & Averianov, A. O. An aegialodontid upper molar and the evolution of mammal dentition. *Science* **313,** 1092 (2006).
75. Chow, M. & Rich, T. H. *Shuotherium dongi*, n. gen. and sp., a therian with pseudo-tribosphenic molars from the Jurassic of Sichuan, China. *Austral. Mamm.* **5,** 127–142 (1982).
76. Sigogneau-Russell, D. Discovery of a Late Jurassic Chinese mammal in the upper Bathonian of England. *C. R. Acad. Sci. II* **327,** 571–576 (1998).
77. Wang, Y.-Q. *et al.* A probable pseudo-tribosphenic upper molar from the Late Jurassic of China and the early radiation of the Holotheria. *J. Vertebr. Paleontol.* **18,** 777–787 (1998).
78. Luo, Z. X. *et al.* Convergent dental evolution in pseudotribosphenic and tribosphenic mammals. *Nature* **450,** 93–97 (2007).
79. Rich, T. H. *et al.* A tribosphenic mammal from the Mesozoic of Australia. *Science* **278,** 1438–1442 (1997).
80. Rich, T. H. *et al.* An advanced ausktribosphenid from the Early Cretaceous of Australia. *Rec. Queen Victoria Mus.* **110,** 1–9 (2001).
81. Flynn, J. J. *et al.* A Middle Jurassic mammal from Madagascar. *Nature* **401,** 57–60 (1999).
82. Rauhut, O. W. M. *et al.* A Jurassic mammal from South America. *Nature* **416,** 165–168 (2002).
83. Martin, T. & Rauhut, O. W. M. Mandible and dentition of *Asfaltomylos patagonicus* (Australosphenida, Mammalia) and the evolution of tribosphenic teeth. *J. Vertebr. Paleontol.* **25,** 414–425 (2005).
84. Rougier, G. W. *et al.* New Jurassic mammals from Patagonia, Argentina: a reappraisal of australosphenidan morphology and interrelationship. *Am. Mus. Novitates* **3566,** 1–54 (2007).
85. Luo, Z.-X. *et al.* Dual origin of tribosphenic mammals. *Nature* **409,** 53–57 (2001).
86. Rich, T. H. *et al.* Evidence that monotremes and ausktribosphenids are not sistergroups. *J. Vertebr. Paleontol.* **22,** 466–469 (2002).
87. Woodburne, M. O. Monotremes as pretribosphenic mammals. *J. Mamm. Evol.* **10,** 195–248 (2003).
88. Musser, A. M. *Investigations into the Evolution of Australian Mammals with a Focus on Monotremata* PhD thesis, Univ. of New South Wales (2006).
89. Sigogneau-Russell, D. *et al.* The oldest tribosphenic mammal from Laurasia (Purbeck Limestone Group, Berriasian, Cretaceous, UK) and its bearing on the "dual origin" of Tribosphenida. *Comptes. Rend. Acad. Sci.* **333,** 141–147 (2001).
90. Sigogneau-Russell, D. Docodonts from the British Mesozoic. *Acta Palaeontol. Polonica* **48,** 357–374 (2003).
91. Evans, A. R. & Sanson, G. D. The tooth of perfection: functional and spatial constraints on mammalian tooth shape. *Biol. J. Linn. Soc.* **78,** 173–191 (2003).
92. Kangas, A. T. *et al.* Nonindependence of mammalian dental characters. *Nature* **432,** 211–214 (2004).
93. Kassai, Y. *et al.* Regulation of mammalian tooth cusp patterning by ectodin. *Science* **309,** 2067–2070 (2005).
94. Jenkins, F. A. Jr. The postcranial skeleton of African cynodonts. *Peabody Mus. Nat. Hist. Bull.* **36,** 1–216 (1971).
95. Filler, A. G. *Axial Character Seriation in Mammals: an Historical and Morphological Exploration of the Origin, Development, Use and Current Collapse of the Homology Paradigm* PhD thesis, Harvard Univ. (1986).
96. Narita, Y. & Kuratani, S. Evolution of vertebral formulae in mammals: a perspective on developmental constraints. *J. Exp. Zool.* **304B,** 91–106 (2005).
97. Wellik, D. M. & Capecchi, M. R. *Hox10* and *Hox11* genes are required to globally pattern the mammalian skeleton. *Science* **301,** 363–367 (2003).
98. Li, G. & Luo, Z.-X. A Cretaceous symmetrodont therian with some monotreme-like postcranial features. *Nature* **439,** 195–200 (2006).

Acknowledgements I benefited from years of stimulating discussion about early mammal evolution with R. Cifelli, T. Martin, J. Wible, Z. Kielan-Jaworowska, T. Rowe, H. Sues, M. Dawson, K. C. Beard, G. Wilson, G. Rougier, J. Bonaparte, W. Maier, P.-J. Chen and Q. Ji, and discussion on diversification pattern with D. Erwin and M. Benton. Many helped my research: A. Tabrum, X.-N. Yang, Q. Yang, P.-J. Chen, Z.-M. Dong, K.-Q. Gao. I thank Q. Ji and J. R. Wible for access to comparative collections; M. R. Dawson, T. Martin and J. R. Wible for improving the manuscript; M. Klingler for assistance with graphics. Support was from the National Science Foundation (USA), National Natural Science Foundation of China, National Geographic Society and the Carnegie Museum.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to Z.-X.L. (luoz@carnegiemnh.org).

ARTICLES

Two stellar components in the halo of the Milky Way

Daniela Carollo[1,2,3,5], Timothy C. Beers[2,3], Young Sun Lee[2,3], Masashi Chiba[4], John E. Norris[5], Ronald Wilhelm[6], Thirupathi Sivarani[2,3], Brian Marsteller[2,3], Jeffrey A. Munn[7], Coryn A. L. Bailer-Jones[8], Paola Re Fiorentin[8,9] & Donald G. York[10,11]

The halo of the Milky Way provides unique elemental abundance and kinematic information on the first objects to form in the Universe, and this information can be used to tightly constrain models of galaxy formation and evolution. Although the halo was once considered a single component, evidence for its dichotomy has slowly emerged in recent years from inspection of small samples of halo objects. Here we show that the halo is indeed clearly divisible into two broadly overlapping structural components—an inner and an outer halo—that exhibit different spatial density profiles, stellar orbits and stellar metallicities (abundances of elements heavier than helium). The inner halo has a modest net prograde rotation, whereas the outer halo exhibits a net retrograde rotation and a peak metallicity one-third that of the inner halo. These properties indicate that the individual halo components probably formed in fundamentally different ways, through successive dissipational (inner) and dissipationless (outer) mergers and tidal disruption of proto-Galactic clumps.

Astronomers have long sought to constrain models for the formation and evolution of the Milky Way (our Galaxy) on the basis of observations of the stellar and globular cluster populations that it contains. These populations are traditionally defined as samples of objects that exhibit common spatial distributions, kinematics and metallicities (the age of a population, when available, is also sometimes used). Metallicity is taken by astronomers to represent the abundances of elements heavier than helium, which are only created by nucleosynthesis in stars—either internally via nuclear burning in their cores or externally during explosive nucleosynthesis at the end of their lives. The earliest generations of stars have the lowest metallicities, because the gas from which they formed had not been enriched in heavy elements created by previous stars and distributed throughout the primordial interstellar medium by stellar winds and supernovae.

Previous work has provided evidence that the halo of the Milky Way may not comprise a single population, primarily from analysis of the spatial profiles (or inferred spatial profiles) of halo objects[1–4]. A recent example of such an analysis is the observation of two different spatial density profiles for distinct classes of RR Lyrae variable stars in the halo[5]. In addition, tentative claims for a net retrograde motion of halo objects by previous authors supports the existence of a likely dual-component halo[6–10]. The central difficulty in establishing with confidence whether or not a dichotomy of the halo populations exists is that the past samples of tracer objects have been quite small, and usually suitable only for consideration of a limited number of the expected signatures of its presence.

In the present work, we examine this question in detail using a large, homogeneously selected and analysed sample of over 20,000 stars, originally obtained as calibration data during the course of the Sloan Digital Sky Survey (SDSS)[11]. Although there are many possible alternative (and more complex) models that might be considered, multiple lines of evidence derived from these data clearly confirm that the halo can be resolved into (at least) two primary populations, the inner and the outer halo, with very different observed properties.

We find that the inner-halo component of the Milky Way dominates the population of halo stars found at distances up to 10–15 kpc from the Galactic Centre (including the solar neighbourhood). An outer-halo component dominates in the regions beyond 15–20 kpc. We show the inner halo to be a population of stars that are non-spherically distributed about the centre of the Galaxy, with an inferred axial ratio of the order of $\sim$0.6. Inner-halo stars possess generally high orbital eccentricities, and exhibit a modest prograde rotation (between 0 and 50 km s^{-1}) around the centre of the Galaxy (see Supplementary Table 1). The distribution of metallicities for stars in the inner halo peaks at [Fe/H] = -1.6, with tails extending to higher and lower metallicities. (Here metallicity is defined as $[A/B] = \log_{10}(N_A/N_B) - \log_{10}(N_A/N_B)_{\odot}$, where N_A and N_B represent the number density of atoms of elements A and B, and the subscript $\odot$ indicates solar values.) The outer halo, by contrast, comprises stars that exhibit a much more spherical spatial distribution, with an axial ratio of $\sim$0.9–1.0. Outer-halo stars cover a wide range of orbital eccentricities, including many with lower eccentricity orbits than found for most stars associated with the inner halo, and exhibit a clear retrograde net rotation (between –40 and –70 km s^{-1}) about the centre of the Galaxy. The metallicity distribution function (MDF) of the outer halo peaks at lower metallicity than that of the inner halo, around [Fe/H] = -2.2, and includes a larger fraction of low-metallicity stars than does the MDF of the inner-halo population.

Evidence for the dichotomy of the halo

The spectroscopy, photometry and astrometry for our large sample of stars were obtained from observations carried out with the Apache Point 2.5-m SDSS telescope; these data are publicly available as Data

[1]INAF–Osservatorio Astronomico di Torino, 10025 Pino Torinese, Italy. [2]Department of Physics and Astronomy, Center for the Study of Cosmic Evolution, [3]Joint Institute for Nuclear Astrophysics, Michigan State University, E. Lansing, Michigan 48824, USA. [4]Astronomical Institute, Tohoku University, Sendai 980-8578, Japan. [5]Research School of Astronomy and Astrophysics, The Australian National University, Mount Stromlo Observatory, Cotter Road, Weston, Australian Capital Territory 2611, Australia. [6]Department of Physics, Texas Tech University, Lubbock, Texas 79409, USA. [7]US Naval Observatory, PO Box 1149, Flagstaff, Arizona 86002, USA. [8]Max-Planck-Institut für Astronomy, Königstuhl 17, D-69117, Heidelberg, Germany. [9]Department of Physics, University of Ljubljana, Jadranska 19, 1000, Ljubljana, Slovenia. [10]Department of Astronomy and Astrophysics, [11]The Enrico Fermi Institute, University of Chicago, Chicago, Illinois 60637, USA.

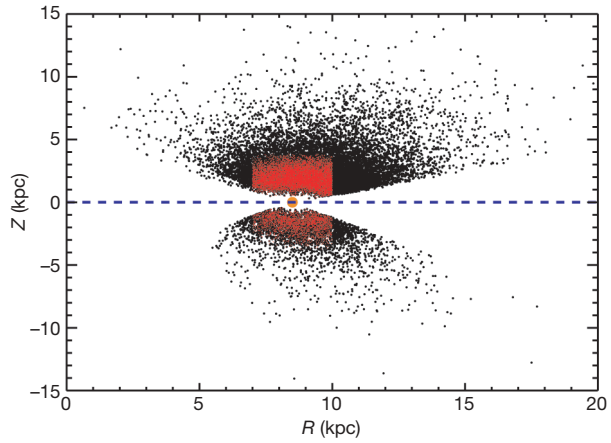


Figure 1 | The spatial distribution of the stars analysed in the present sample. The distribution of the full sample of 20,236 unique SDSS (Data Release 5[12]) spectrophotometric and telluric calibration stars in the Z–R plane is shown, where Z is the derived distance from the Galactic plane in the vertical direction and R is the derived distance from the centre of the Galaxy projected onto this plane. The dashed blue line represents the Galactic plane, while the filled orange dot is the position of the Sun, at $Z = 0$ kpc and $R = 8.5$ kpc. The 'wedge shape' of the selection area is the result of limits of the SDSS footprint in Galactic latitude. The red points indicate the 10,123 stars that satisfy our criteria for a local sample of stars, having 7 kpc $< R <$ 10 kpc, with distance estimates from the Sun $d < 4$ kpc, and with viable measurements of stellar parameters and proper motions.

Release 5[12]. Details concerning the selection of the stars and the measurement of their parameters (temperature, surface gravity and metallicity, [Fe/H]), as well as the methods used to obtain their estimated distances, proper motions and derived kinematics, can be found in the Supplementary Information.

Our 20,236 programme stars explore distances up to 20 kpc from the Sun, but we can only obtain useful estimates of the full space motions (as described in the Supplementary Information) for the subset of 10,123 stars in a local volume (up to 4 kpc from the Sun; Fig. 1). The restriction of the sample to the region of the solar neighbourhood is also made so that the assumptions going into the kinematic calculations are best satisfied. Figure 2 shows the distribution of [Fe/H] for different cuts in the V velocity, which is the orbital component that measures the motion of a star (with respect to the Local Standard of Rest) in the rotation direction of the Galaxy. The transition in the distribution of [Fe/H] that is expected as one sweeps from stars with thick-disk-like motions to stars with halo-like motions is clear. However, with the large sample of stars in our sample, it is possible to investigate the change in the distribution of [Fe/H] for stars that are increasingly more retrograde, as well as for those that are both highly retrograde and have orbits taking them to high Z_{max} (the maximum distance above the Galactic plane reached by a star during the course of its orbit about the Galactic Centre—the Supplementary Information describes the methods used to derive this fundamental parameter). This figure shows that stars with the most retrograde orbits, and those that reach large distances in their orbits above the Galactic plane, exhibit distributions of [Fe/H] that peak at metallicities between −2.0 and −2.2, which we associate with the outer-halo population. The inner-halo population dominates the samples of stars with peak metallicity [Fe/H] ≈ −1.6.

Astronomers have long debated whether there might exist a change in the rotational properties of the halo of the Milky Way as a function of distance from the Galactic Centre, based on much smaller samples of globular clusters[2,10] and stars[6–9,13,14] than we consider here. The

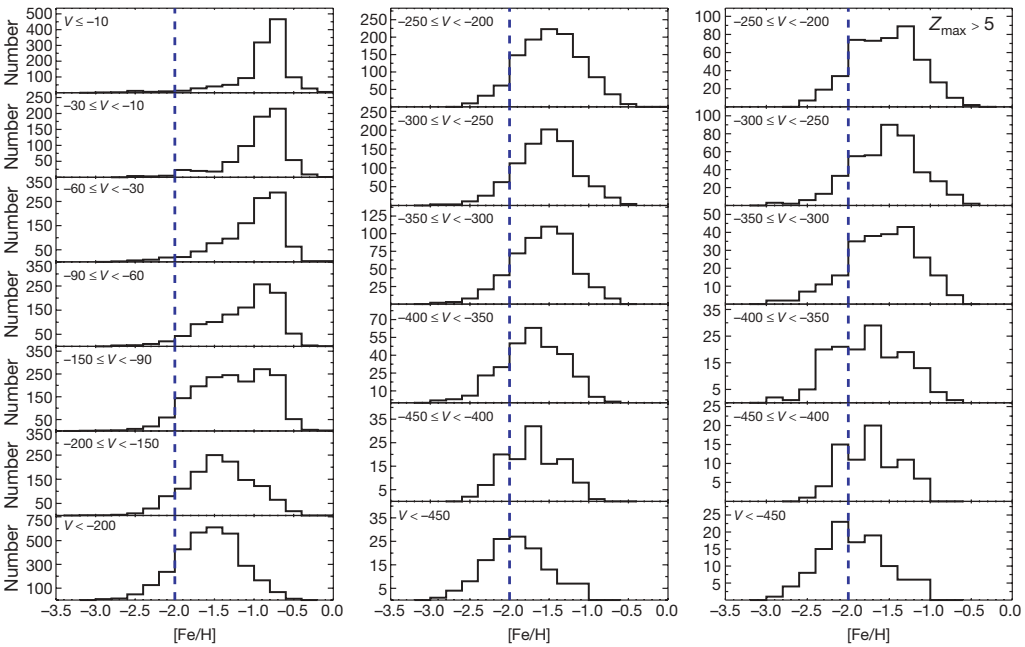


Figure 2 | The distribution of [Fe/H] for various cuts in the V velocity (in km s^{-1}), the component of orbital motion measured with respect to the Local Standard of Rest. The Local Standard of Rest is a frame in which the mean space motions of the stars in the solar neighbourhood average to zero. A blue dashed line at [Fe/H] = −2.0 is added for reference in all three columns. In the left-hand column, the full data set is considered. The stars with modestly negative V velocities in the upper three panels are dominated by stars from the thick-disk (and metal-weak thick-disk) populations, with a peak metallicity around [Fe/H] ≈ −0.7. A transition to dominance by inner- and outer-halo population stars becomes evident for $V < -90$ km s^{-1}; in the bottom panel of this column, the distribution of [Fe/H] appears similar to what in the past was considered 'the halo', but we argue results from a superposition of contributions from both inner- and outer-halo populations. In the middle column, the large numbers of stars with $V < -200$ km s^{-1} are broken into smaller ranges in V velocity. As V becomes increasingly retrograde ($V < -220$ km s^{-1}), the metallicity distribution shifts to include ever larger numbers of stars with [Fe/H] < −2.0, and relatively fewer stars with [Fe/H] ≈ −1.6. The same V velocity cuts are applied in the right-hand column, but only for stars with $Z_{max} > 5$ kpc, in order to decrease the contribution from inner-halo stars. Although fewer stars are included, the increasing dominance of stars with [Fe/H] < −2.0 is even more apparent. We associate the stars with the most extreme retrograde orbits (and those that reach far above the Galactic plane in their orbits) with the outer-halo population.

stellar samples were obtained with selection criteria (for example, on the basis of high proper motions for halo stars in the solar neighbourhood[7,13,14], or from *in situ*, apparent-magnitude limited surveys[6,9,15]) that we suggest favoured membership in one or the other of the now clearly revealed halo components. A detailed summary of the kinematics of our programme stars is presented in the Supplementary Information, where we also establish consistency between properties obtained through techniques based on full space motions and those based on radial velocities alone, which argues against the existence of any large systematic errors in the proper motions.

Table 1 summarizes the past and present determinations of $\langle V_\Phi \rangle$, the mean rotational velocity with respect to the Galactic Centre, where claims for a retrograde halo have been made. Previous samples that have addressed this question were based on either much smaller total numbers of objects (with the limitation that they could not well sample both the inner- and outer-halo populations), did not have proper motions available (or only highly uncertain ones), or were otherwise restricted due to the selection criteria employed (that is, they were kinematically biased[16], or had limited sky coverage, rendering them sensitive to the effects of individual star streams[17]). The local sample of SDSS calibration stars we have assembled does not suffer from any of these limitations. The retrograde signatures for stars we associate with the outer-halo population are robust and highly statistically significant (except for the smallest subsample). However, even our precise present determination of the net retrograde rotation of the outer halo, based on our local sample, is probably influenced by some degree of overlap between outer-halo stars with those from the inner-halo population.

The distribution of [Fe/H] for stars on increasingly retrograde orbits about the Galactic Centre for subsamples that reach different distances from the Galactic plane (Z_{max}) is shown in Fig. 3. The MDFs of the stars with Z_{max} close to the Galactic plane are very different from those whose orbits reach farther from the plane. The distribution of metallicity clearly shifts to lower abundances as more severe cuts on V_Φ or Z_{max} are applied, as supported by rigorous statistical tests. We conclude that the halo of the Galaxy comprises stars with intrinsically different distributions of [Fe/H]; the observed changes in the MDF of halo stars with V_Φ and Z_{max} would not be expected if the halo is considered as a single entity. The Supplementary Information presents additional observed differences in the energetics, the distribution of orbital eccentricities, and changes in the nature of stellar orbits for our programme stars that are also inconsistent with a single halo population.

In order to provide confirmation of the shift in the MDF inferred from our analysis of a local sample of stars, we also examine an auxiliary sample of stars that are at present located much farther from the Galactic Centre. This sample comprises 1,235 blue horizontal-branch stars selected from the SDSS[18]. The stars cover a wide range of distances, from 5 kpc to over 80 kpc from the centre of the Galaxy. Statistical tests strongly reject the hypothesis that the stars at large distances from the Galactic Centre could be drawn from the same parent population as those at distances close to the Galactic Centre (Fig. 4).

It has been shown, on the basis of Jeans' theorem[19,20], that the global structure of the stellar halo can be recovered from local kinematic information, as long as one has a sufficiently large number of stars observed in the solar neighbourhood that explore the full phase-space distribution of the pertinent stellar populations. We note that the actual halo systems of the Galaxy are unlikely to be in well-mixed equilibrium states. However, the relaxation process is very slow compared to the orbital periods of typical stars, so Jeans' theorem and the approach based on it remain at least approximately valid. The result of this exercise for our large sample of SDSS calibration stars, over narrow cuts in metallicity, is shown in Fig. 5. The observed changes in the inferred spatial density profiles suggest that a flattened inner-halo population dominates locally for stars with [Fe/H] > -2, whereas the outer-halo population has a nearly spherical distribution, and dominates at distances beyond $r \approx 15$–20 kpc (where r represents the distance from the Galactic Centre), as well as locally for stars with [Fe/H] < -2.0. Variations in the halo spatial profile with distance have been recognized by a number of previous authors[1–4,15,20], based on samples of stars that are one to two orders of magnitude smaller than our present data set.

Implications of the dichotomy of the halo

An early model for the formation of the Milky Way, based on the rapid (a few hundred million years) monolithic collapse of a gaseous proto-Galaxy[21], has yielded to the more recent idea that the halo of

Table 1 | Studies claiming a retrograde outer halo

Sample and selection criteria	N	Additional restrictions	$\langle V_\Phi \rangle$ (km s^{-1})	Method	Source
Globular clusters (non-kinematic)	19	'Young halo'	-64 ± 74	F&W	Ref. 2
Globular clusters (non-kinematic)	20	'Young halo'	-42 ± 80	F&W	Ref. 10
RR Lyrae stars (non-kinematic)	26	$\lvert Z \rvert < 8$ kpc	-95 ± 29	FSM	Ref. 9
Field subdwarfs (kinematic)	30	$Z_{max} > 5$ kpc	-45 ± 22	FSM	Ref. 7
		Bias corrected	$+24 \pm 13$		Ref. 16
Field horizontal-branch stars (non-kinematic)	90	[Fe/H] < -1.6	-93 ± 36	F&W	Ref. 8
		$\lvert Z \rvert > 4$ kpc			
Field subdwarfs (kinematic)	101	$V < -100$ km s^{-1}	-32 ± 10	FSM	Ref. 13
		[Fe/H] < -1.8			
Field F, G, K dwarfs (non-kinematic)	250	$\lvert Z \rvert > 5$ kpc	-55 ± 16	FSM	Ref. 6
Field F, G turnoff (non-kinematic)	2,228	$Z_{max} > 5$ kpc	-11 ± 2	FSM	This work
	200	[Fe/H] < -1.0	-41 ± 11		
		[Fe/H] < -2.2			
	771	$Z_{max} > 10$ kpc	-38 ± 5		
	94	[Fe/H] < -1.0	-71 ± 17		
		[Fe/H] < -2.2			
	371	$Z_{max} > 15$ kpc	-56 ± 8		
	54	[Fe/H] < -1.0	-71 ± 25		
		[Fe/H] < -2.2			

Previous and current determinations of the mean rotational velocity, $\langle V_\Phi \rangle$, and the error in the mean ($\sigma/\sqrt{N}$, where σ is the standard deviation and N is the number of stars), for samples in which a counter-rotating halo has been claimed, ordered by sample size. The samples listed in the first column are classified as to whether they were chosen on the basis of high proper motions (kinematic) or not (non-kinematic). Restrictions placed on each sample by the authors are listed in the third column (see original papers for details). The method of analysis used for each determination is listed: F&W, estimate based on the technique of Frenk and White[48], which considers distances and radial velocities alone, under the assumption of a cylindrically symmetric Galaxy; FSM, estimate based on consideration of the full space motions, which requires the use of proper motions, as well as distances and radial velocities. The samples analysed with the F&W approach are either not statistically different from zero (refs 2, 10), or are only marginally so (2.6σ; ref. 8). Previous samples based on analysis of the full space motions vary from statistically insignificant (ref. 7), to just over 3σ significance (refs 6, 9, 13), owing to the small numbers of stars considered. Note that after application of (uncertain) corrections for kinematic bias (ref. 16), the retrograde result reported in ref. 7 disappears entirely. One can assume that a similar outcome might apply to the ref. 13 determination. The samples of ref. 6 and ref. 9 are both selected over a restricted region of sky (towards the North Galactic Pole) and are therefore subject to possible contamination by individual stellar streams. All but two subsamples of the SDSS calibration star sample have retrograde signals that are significant at more than the 4σ level. The subsample at $Z_{max} > 5$ kpc and [Fe/H] < -1 is likely to include significant contamination from inner-halo stars.

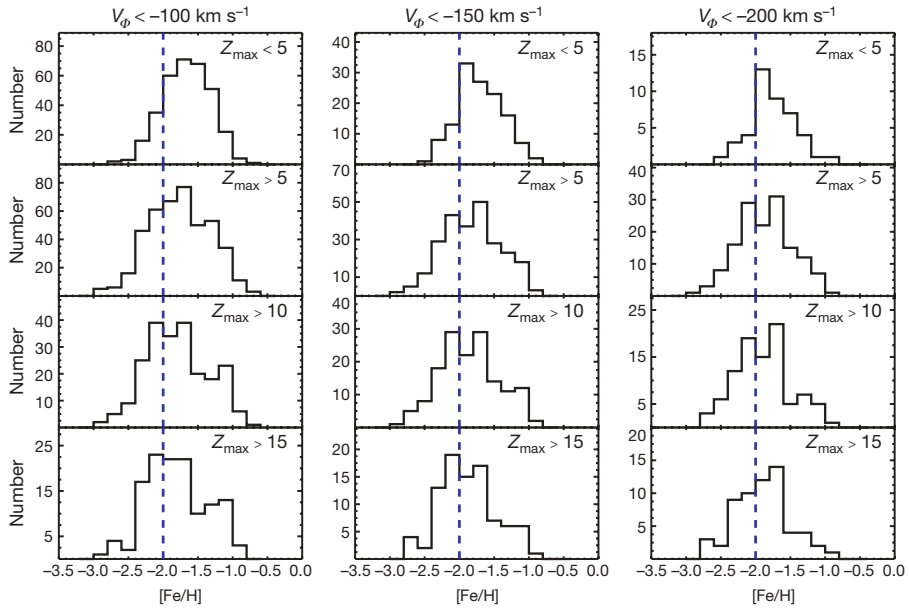


Figure 3 | The distribution of [Fe/H] for the stars in our sample on highly retrograde orbits. Stars from the disk populations, which possess prograde orbits, cannot be present in this plot. The panels show various cuts in V_Φ, the rotational velocity with respect to the Galactic Centre in a cylindrical coordinate system, and for different ranges of Z_{max} (in kpc). A blue dashed line at [Fe/H] = −2.0 is added for reference. The left-hand column applies for stars with $V_\Phi < -100$ km s^{-1}; the clearly skewed distribution of [Fe/H] exhibits an increased contribution from lower metallicity stars as one progresses from the low ($Z_{max} < 5$ kpc) to the high ($Z_{max} > 15$ kpc) subsamples. Simultaneously, the predominance of stars from the inner-halo population, with peak metallicity at [Fe/H] ≈ −1.6, decreases in relative strength, and shifts to lower [Fe/H]. Similar behaviours are seen in the middle and right-hand columns, which correspond to cuts on $V_\Phi < -150$ km s^{-1} and -200 km s^{-1}, respectively. A Kolmogorov-Smirnoff test of the null hypothesis that the MDFs of stars shown in the lower panels for the individual cuts on V_Φ could be drawn from the MDFs of the same parent population as those shown in the upper panels, against an alternative that the stars are drawn from more metal-poor parent MDFs, is rejected at high levels of statistical significance. For $V_\Phi < -100$ km s^{-1}, one-sided probabilities less than 0.0001 are obtained for the cuts on $Z_{max} > 5$, 10 and 15 kpc. For $V_\Phi < -150$ km s^{-1}, one-sided probabilities of 0.0004, 0.0001 and 0.0003 are obtained for $Z_{max} > 5$, 10 and 15 kpc, respectively. For $V_\Phi < -200$ km s^{-1}, one-sided probabilities of 0.014, 0.010 and 0.033 are obtained, for $Z_{max} > 5$, 10 and 15 kpc, respectively.

the Galaxy was assembled, over the span of several billion years, from smaller proto-Galactic clumps[22]. This hierarchical assembly model has received close attention in recent years, in part because it fits well with the prevailing theory for the formation and evolution of structure in the Universe, based on the early collapse of 'mini-haloes' of cold dark matter (CDM)[23,24]. Modern numerical simulations for the assembly of large spirals based on CDM cosmogonies predict that the stars in the haloes of galaxies like the Milky Way might be composed of the shredded stellar debris of numerous dwarf-like galaxies that have been torn apart by tidal interactions with their parent galaxy[24–27]. Recent quantitative analysis of the amount of structure visible in the halo of the Galaxy from SDSS[28–31] imaging provides compelling additional evidence[32]. Others have argued that some combination of a monolithic collapse and a hierarchical assembly model may be necessary to fully explain the observed data[2,15,33].

Within the context of the CDM model, the formation of the inner halo may be understood in the following manner. Low-mass sub-Galactic fragments are formed at an early stage. These fragments

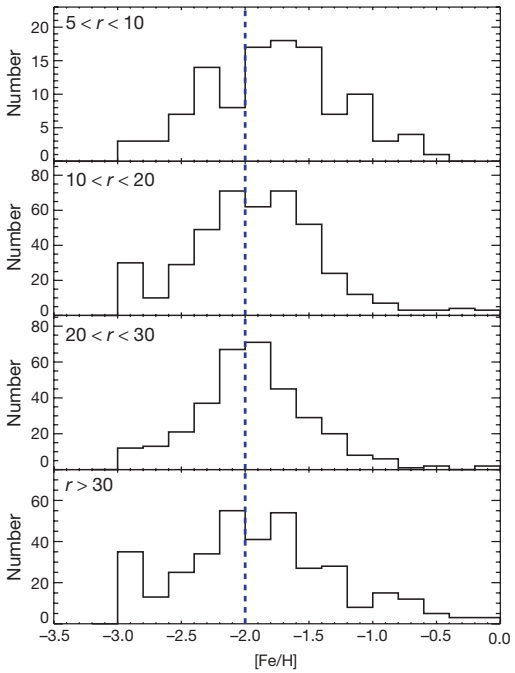


Figure 4 | A sample of blue horizontal-branch stars exploring much larger distances from the Galactic Centre than the SDSS calibration stars. The distribution of [Fe/H] is shown for various cuts on the distance from the Galactic Centre, *r*, in kpc. The nature of the MDF appears to shift from the upper two panels, which exhibit the character of a mixture of inner- and outer-halo populations, over to a unimodal distribution in the third panel, centred on [Fe/H] ≈ −2.0. The most distant blue horizontal branch (BHB) stars in the lowest panel also exhibit the appearance of a mixture of the two populations, possibly due to the inclusion of inner-halo stars on highly eccentric orbits that take them far from the Galactic Centre. The peak around [Fe/H] = −3.0 seen in several of the panels is an artefact arising from the limit of the metallicity grid that is used for abundance determinations of the BHB stars. A Kolmogorov-Smirnoff test of the null hypothesis that the MDFs of stars shown in the lower panels for the individual cuts on Galactocentric distance *r* could be drawn from the same parent population as the stars shown in the first panel, against an alternative that the stars are drawn from more metal-poor parent MDFs, is rejected at high levels of statistical significance (one-sided probabilities of 0.0262, 0.0005 and 0.0243, respectively, for the three higher cuts on *r*). The fraction of stars with metallicities [Fe/H] < −2.0 (primarily outer-halo stars) grows from 31% for stars with 5 < *r* < 10 kpc to 46% for stars at larger Galactocentric distances.

rapidly merge into several (in many simulations, two[26,34]) more-massive clumps, which themselves eventually dissipatively merge (owing to the presence of gas that has yet to form stars). The essentially radial merger of the few resulting massive clumps gives rise to the dominance of the high-eccentricity orbits for stars that we assign here to membership in the inner halo. Star formation within these massive clumps (both pre- and post-merger) would drive the mean metallicity to higher abundances. This is followed by a stage of adiabatic compression (flattening) of the inner halo component owing to the growth of a massive disk, along with the continued accretion of gas onto the Galaxy[34,35].

The fact that the outer-halo component of the Milky Way exhibits a net retrograde rotation (and a different distribution of overall orbital properties), as found here, clearly indicates that the formation of the outer halo is distinct from that of both the inner-halo and disk components. We suggest, as others have before, that the outer-halo component formed, not through a dissipative, angular-momentum-conserving contraction, but rather through dissipationless chaotic merging of smaller subsystems within a pre-existing dark-matter halo. These subsystems would be expected to be of much lower mass, and subject to tidal disruption in the outer part of a dark-matter halo, before they fall farther into the inner part. As candidate (surviving) counterparts for such subsystems, one might consider the low-luminosity dwarf spheroidal galaxies surrounding the Galaxy, in particular the most extreme cases recently identified from the SDSS[36,37]. Subsystems of lower mass, and by inference, even lower metallicity, may indeed be destroyed so effectively that none (or very few) have survived to the present day. If so, the outer-halo population may be assembled from relatively more metal-poor stars, following the luminosity–metallicity relationship for Local Group dwarf galaxies[38].

The net retrograde rotation of the outer halo may be understood in the context of the higher efficiency of phase mixing for the orbits of stars that are stripped from subsystems on prograde, rather than retrograde, orbits[39,40].

The clear difference in the MDFs of the two halo populations we identify also suggests that the lowest metallicity stars in the Galaxy may be associated with the outer halo, which can be exploited for future directed searches. It is noteworthy that the hyper metal-poor stars HE 0107−5240[41] and HE 1327−2326[42], both of which have [Fe/H] < -5.0, as well as the recently discovered ultra metal-poor star HE 0557−4840[43], with [Fe/H] $= -4.8$, are either located greater than 10 kpc away (HE 0107−5240, HE 0557−4840) or have space motions that carry them far out into the Galaxy (HE 1327−2326; A. Frebel, personal communication).

In addition, efforts to determine the primordial lithium abundance from observations of the most metal-poor stars[44] may have inadvertently mixed samples from the inner- and outer-halo populations; such stars could have formed and evolved in rather different astrophysical environments. The inner/outer halo dichotomy may also have an impact on the expected numbers of carbon-enhanced metal-poor stars as a function of declining metallicity[45], and as a function of distance from the Galactic plane[46,47].

Much remains to be learned as the database of low-metallicity stars continues to expand, in particular from those stars that are found in distant *in situ* samples, or from those nearby stars with available proper motions that indicate membership of the outer-halo population. We look forward to the next dramatic increase in the numbers of very metal-poor stars that will come from the ongoing stellar samples from SDSS, and in particular from SEGUE, the Sloan Extension for Galactic Understanding and Exploration.

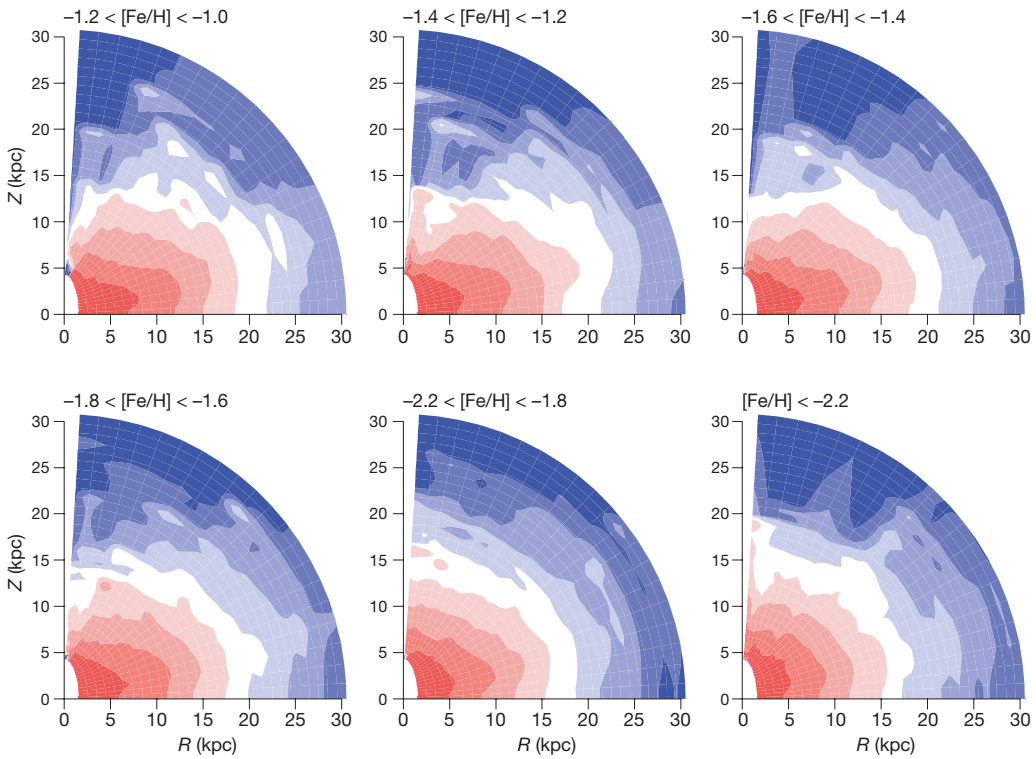


Figure 5 | Equidensity contours of the reconstructed global density distributions for stars in our sample with various metallicities. The global density distributions are constructed from the sum of the probability density of an orbit at each location in the Z–R plane, with a weighting factor being inversely proportional to the corresponding density at the currently observed position of the star[19,20]. High-density regions are indicated by redder colours, while low-density regions are indicated by bluer colours (a linear density scale is used). Within each metallicity cut, the apparent flattening of the inner regions slowly goes over to a more spherical shape with increasing distance. As one progresses from the more metal-rich ([Fe/H] ≈ -1.0) to the most metal-poor ([Fe/H] < -2.2) subsets of these data, the overall nature of the equidensity contours also changes from highly flattened (axial ratios of ~0.6), to more spherical (axial ratio of ~0.9). This suggests that the inner- and outer-halo components are broadly overlapping in space and in metallicity—the inner-halo population is characterized as a flattened density distribution that dominates locally for stars with [Fe/H] > -2, whereas the outer-halo population is nearly spherical, and dominates at larger distances and locally for stars with [Fe/H] < -2.

Received 20 June; accepted 5 November 2007.

1. Hartwick, F. D. A. in *The Galaxy* (eds Gilmore, G. & Carswell, B.) 281–290 (NATO ASI Series 207, Reidel, Dordrecht, 1987).
2. Zinn, R. in *The Globular Clusters-Galaxy Connection* (eds Smith, G. H. & Brodie, J. P.) 38–47 (ASP Conf. Ser. 48, Astronomical Society of the Pacific, San Francisco, 1993).
3. Preston, G. W., Shectman, S. A. & Beers, T. C. Detection of a galactic color gradient for blue horizontal-branch stars of the halo field and implications for the halo age and density distributions. *Astrophys. J.* **375,** 121–147 (1991).
4. Kinman, T. D., Suntzeff, N. B. & Kraft, R. P. The structure of the galactic halo outside the solar circle as traced by the blue horizontal branch stars. *Astron. J.* **108,** 1722–1772 (1994).
5. Miceli, A. *et al.* Evidence for distinct components of the Galactic stellar halo from 838 RR Lyrae stars discovered in the LONEOS-I survey. *Astrophys. J.* (in the press); preprint at ⟨http://arxiv.org/abs/0706.1583⟩ (2007).
6. Majewski, S. R. A complete, multicolor survey of absolute proper motions to B of about 22.5 – Galactic structure and kinematics at the north Galactic pole. *Astrophys. J.* **78** (Suppl.), 87–152 (1992).
7. Carney, B. W., Laird, J. B., Latham, D. W. & Aguilar, L. A. A survey of proper motion stars. XIII. The halo population(s). *Astron. J.* **112,** 668–692 (1996).
8. Wilhelm, R. *et al.* in *Formation of the Galactic Halo... Inside and Out* (eds Morrison, H. & Sarajedini, A.) 171–174 (ASP Conf. Ser. 92, Astronomical Society of the Pacific, San Francisco, 1996).
9. Kinman, T. D., Cacciari, C., Bragaglia, A., Buzzoni, A. & Spagna, A. Kinematic structure in the Galactic halo at the north Galactic pole: RR Lyrae and BHB stars show different kinematics. *Mon. Not. R. Astron. Soc.* **371,** 1381–1398 (2007).
10. Lee, Y.-W., Hansung, B. G. & Casetti-Dinescu, D. I. Kinematic decoupling of globular clusters with extended horizontal branches. *Astrophys. J.* **661,** L49–L52 (2007).
11. York, D. G. *et al.* The Sloan Digital Sky Survey: Technical summary. *Astron. J.* **120,** 1579–1587 (2000).
12. Adelman-McCarthy, J. K. *et al.* The fifth data release of the Sloan Digital Sky Survey. *Astrophys. J.*, Suppl. 172, 634–644 (2007).
13. Sandage, A. & Fouts, G. New subdwarfs. VI. Kinematics of 1125 high-proper-motion stars and the collapse of the Galaxy. *Astron. J.* **92,** 74–115 (1987).
14. Ryan, S. G. & Norris, J. E. Subdwarf studies. II – Abundances and kinematics from medium-resolution spectra. III. – The halo metallicity distribution. *Astron. J.* **101,** 1835–1864 (1991).
15. Chiba, M. & Beers, T. C. Kinematics of metal-poor stars in the Galaxy. III. Formation of the stellar halo and thick disk as revealed from a large sample of non-kinematically selected stars. *Astron. J.* **119,** 2843–2865 (2000).
16. Carney, B. W. in *The Third Stromlo Symposium: The Galactic Halo* (eds Gibson, B. K., Axelrod, T. S. & Putman, M. E.) 230–242 (ASP Conf. Ser. 165, Astronomical Society of the Pacific, San Francisco, 1999).
17. Majewski, S. R., Munn, J. A. & Hawley, S. L. Absolute proper motions to B approximately 22.5: Evidence for kimematical substructure in halo field stars. *Astrophys. J.* **427,** L37–L41 (1994).
18. Sirko, E. *et al.* Blue horizontal-branch stars in the Sloan Digital Sky Survey. I. Sample selection and structure in the Galactic halo. *Astron. J.* **127,** 899–913 (2004).
19. Binney, J. & May, A. The spheroids of galaxies before and after disc formation. *Mon. Not. R. Astron. Soc.* **218,** 743–760 (1986).
20. Sommer-Larsen, J. & Zhen, C. Armchair cartography – A map of the Galactic halo based on observations of local, metal-poor stars. *Mon. Not. R. Astron. Soc.* **242,** 10–24 (1990).
21. Eggen, O. J., Lynden-Bell, D. & Sandage, A. R. Evidence from the motions of old stars that the galaxy collapsed. *Astrophys. J.* **136,** 748–766 (1962).
22. Searle, L. & Zinn, R. Compositions of halo clusters and the formation of the galactic halo. *Astrophys. J.* **225,** 357–379 (1978).
23. White, S. D. M. & Rees, M. J. Core condensation in heavy halos – A two-stage theory for galaxy formation and clustering. *Mon. Not. R. Astron. Soc.* **183,** 341–358 (1978).
24. Moore, B., Diemand, J., Madau, P., Zemp, M. & Stadel, J. Globular clusters, satellite galaxies and stellar haloes from early dark matter peaks. *Mon. Not. R. Astron. Soc.* **368,** 563–570 (2006).
25. Bullock, J. S. & Johnston, K. V. Tracing galaxy formation with stellar halos. I. Methods. *Astrophys. J.* **635,** 931–949 (2005).
26. Abadi, M. G., Navarro, J. F. & Steinmetz, M. Stars beyond galaxies: The origin of extended luminous haloes around galaxies. *Mon. Not. R. Astron. Soc.* **365,** 747–758 (2006).
27. Brook, C. B., Kawata, D., Martel, H., Gibson, B. K. & Scannapeico, E. Chemical and dynamical properties of the stellar halo. *EAS Publ. Ser.* **24,** 269–275 (2007).
28. Fukigita, M. *et al.* The Sloan Digital Sky Survey photometric system. *Astron. J.* **111,** 1748–1756 (1996).
29. Gunn, J. E. *et al.* The Sloan Digital Sky Survey photometric camera. *Astron. J.* **116,** 3040–3081 (1998).
30. Pier, J. R. *et al.* Astrometric calibration of the Sloan Digital Sky Survey. *Astron. J.* **125,** 1559–1579 (2003).
31. Gunn, J. E. *et al.* The 2.5 m telescope of the Sloan Digital Sky Survey. *Astron. J.* **131,** 2332–2359 (2006).
32. Bell, E. F. *et al.* The accretion origin of the Milky Way's stellar halo. *Astrophys. J* (in the press); preprint at ⟨http://arxiv.org/abs/0706.0004⟩ (2007).
33. Majewski, S. R. Galactic structure surveys and the evolution of the Milky Way. *Annu. Rev. Astron. Astrophys.* **31,** 575–638 (1993).
34. Bekki, K. & Chiba, M. Formation of the galactic stellar halo. I. Structure and kinematics. *Astrophys. J.* **558,** 666–686 (2001).
35. Chiba, M. & Beers, T. C. Structure of the galactic stellar halo prior to disk formation. *Astrophys. J.* **549,** 325–336 (2001).
36. Belokurov, V. *et al.* The field of streams: Sagittarius and its siblings. *Astrophys. J.* **642,** L137–L140 (2006).
37. Belokurov, V. *et al.* Cats and dogs, hair and a hero: A quintet of new Milky Way companions. *Astrophys. J.* **654,** 897–906 (2007).
38. Dekel, A. & Woo, J. Feedback and the fundamental line of low-luminosity low-surface-brightness/dwarf galaxies. *Mon. Not. R. Astron. Soc.* **344,** 1131–1144 (2003).
39. Quinn, P. J. & Goodman, J. Sinking satellites of spiral systems. *Astrophys. J.* **309,** 472–495 (1986).
40. Norris, J. E. & Ryan, S. G. Population studies: Evidence for accretion of the galactic halo. *Astrophys. J.* **336,** L17–L19 (1989).
41. Christlieb, N. *et al.* A stellar relic from the early Galaxy. *Nature* **419,** 904–906 (2002).
42. Frebel, A. *et al.* Nucleosynthetic signatures of the first stars. *Nature* **434,** 871–873 (2005).
43. Norris, J. E. *et al.* HE 0557–4840 – ultra metal-poor and carbon-rich. *Astrophys. J.* **670,** 774–788 (2007).
44. Bonifacio, P. *et al.* First stars VII. Lithium in extremely metal-poor dwarfs. *Astron. Astrophys. J.* **462,** 851–864 (2007).
45. Lucatello, S. *et al.* The frequency of carbon-enhanced metal-poor stars in the Galaxy from the HERES sample. *Astrophys. J.* **653,** L37–L40 (2006).
46. Frebel, A. *et al.* Bright metal-poor stars from the Hamburg/ESO Survey. I. Selection and follow-up observations from 329 fields. *Astrophys. J.* **652,** 1585–1683 (2006).
47. Tumlinson, J. Carbon-enhanced metal-poor stars, the cosmic microwave background, and the stellar IMF in the early universe. *Astrophys. J.*. (submitted).
48. Frenk, C. S. & White, S. D. M. The kinematics and dynamics of the galactic globular cluster system. *Mon. Not. R. Astron. Soc.* **193,** 295–311 (1980).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Allende Preito, E. Bell, W. Brown, A. Frebel, B. Gibson, H. Morrison, C. Thom, J. Tumlinson and B. Yanny for comments on previous versions of this Article. D.C. acknowledges partial support for travel and living expenses from JINA, the Joint Institute for Nuclear Astrophysics, while in residence at Michigan State University. Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the US Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS website is http://www.sdss.org.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to D.C. (carollo@mso.anu.edu.au).

ERRATUM

doi:10.1038/nature06542

Two stellar components in the halo of the Milky Way

Daniela Carollo, Timothy C. Beers, Young Sun Lee, Masashi Chiba, John E. Norris, Ronald Wilhelm, Thirupathi Sivarani, Brian Marsteller, Jeffrey A. Munn, Coryn A. L. Bailer-Jones, Paola Re Fiorentin & Donald G. York

Nature **450**, 1020–1025 (2007)

In Table 1 of this Article, rows 12 to 20 (the 'Field F, G turnoff (non-kinematic)') were inadvertently moved up one row in the N and $\langle V_\phi \rangle$ columns. The corrected table is shown below.

Table 1 | Studies claiming a retrograde outer halo

Sample and selection criteria	N	Additional restrictions	$\langle V_\Phi \rangle$ (km s^{-1})	Method	Source
Globular clusters (non-kinematic)	19	'Young halo'	-64 ± 74	F&W	Ref. 2
Globular clusters (non-kinematic)	20	'Young halo'	-42 ± 80	F&W	Ref. 10
RR Lyrae stars (non-kinematic)	26	$\lvert Z \rvert < 8$ kpc	-95 ± 29	FSM	Ref. 9
Field subdwarfs (kinematic)	30	$Z_{max} > 5$ kpc	-45 ± 22	FSM	Ref. 7
		Bias corrected	$+24 \pm 13$		Ref. 16
Field horizontal-branch stars (non-kinematic)	90	[Fe/H] < -1.6	-93 ± 36	F&W	Ref. 8
		$\lvert Z \rvert > 4$ kpc			
Field subdwarfs (kinematic)	101	$V < -100$ km s^{-1}	-32 ± 10	FSM	Ref. 13
		[Fe/H] < -1.8			
Field F, G, K dwarfs (non-kinematic)	250	$\lvert Z \rvert > 5$ kpc	-55 ± 16	FSM	Ref. 6
Field F, G turnoff (non-kinematic)		$Z_{max} > 5$ kpc		FSM	This work
	2,228	[Fe/H] < -1.0	-11 ± 2		
	200	[Fe/H] < -2.2	-41 ± 11		
		$Z_{max} > 10$ kpc			
	771	[Fe/H] < -1.0	-38 ± 5		
	94	[Fe/H] < -2.2	-71 ± 17		
		$Z_{max} > 15$ kpc			
	371	[Fe/H] < -1.0	-56 ± 8		
	54	[Fe/H] < -2.2	-71 ± 25		

ARTICLES

Molecular code for transmembrane-helix recognition by the Sec61 translocon

Tara Hessa[1]*, Nadja M. Meindl-Beinker[1]*, Andreas Bernsel[2]*, Hyun Kim[1], Yoko Sato[1], Mirjam Lerch-Bader[1], IngMarie Nilsson[1], Stephen H. White[3] & Gunnar von Heijne[1,2]

Transmembrane α-helices in integral membrane proteins are recognized co-translationally and inserted into the membrane of the endoplasmic reticulum by the Sec61 translocon. A full quantitative description of this phenomenon, linking amino acid sequence to membrane insertion efficiency, is still lacking. Here, using *in vitro* translation of a model protein in the presence of dog pancreas rough microsomes to analyse a large number of systematically designed hydrophobic segments, we present a quantitative analysis of the position-dependent contribution of all 20 amino acids to membrane insertion efficiency, as well as of the effects of transmembrane segment length and flanking amino acids. The emerging picture of translocon-mediated transmembrane helix assembly is simple, with the critical sequence characteristics mirroring the physical properties of the lipid bilayer.

Most integral membrane proteins are composed of bundles of tightly packed transmembrane (TM) α-helices[1]. The lipid bilayers into which these proteins are inserted are highly anisotropic and their physicochemical characteristics vary markedly over short distances[2]. This anisotropy is reflected in the distribution of different amino acids in the membrane-embedded parts of integral membrane proteins[3], but the actual recognition of TM helices in nascent polypeptide chains is performed by so-called translocons, complex molecular machines that ensure both the translocation of globular proteins across membranes and the integration of membrane proteins into membranes[4].

What is the 'molecular code' that allows a translocon to recognize TM helices in newly synthesized membrane proteins? We have recently described a system in which ΔG_{app}, the apparent free energy of insertion of a TM helix into the membrane of the endoplasmic reticulum, can be measured and have presented a 'biological' hydrophobicity scale based on such measurements[5]. This scale quantifies the contribution of each of the 20 amino acids, ΔG^{aa}_{app}, to ΔG_{app} for residues placed in the middle position in a 19-residue TM helix. However, if, as has been suggested[5,6], the recognition of TM helices by the Sec61 translocon is based on a thermodynamic partitioning into the anisotropic environment of the lipid bilayer, ΔG^{aa}_{app} should vary with the residue's position in the membrane[7]. ΔG_{app} may also be expected to vary with the overall length of the TM helix, and possibly with the nature of the residues immediately flanking the helix.

To arrive at a full quantitative description of TM helix recognition by the Sec61 translocon, we used the experimental setup summarized in Fig. 1. In brief, systematically designed test segments (H-segments) are introduced near the middle of the large luminal P2 domain of the model protein Lep; the protein is then expressed *in vitro* in the presence of endoplasmic-reticulum-derived dog pancreas rough microsomes, and an apparent equilibrium constant for membrane integration of the H-segment is calculated on the basis of the amount of singly versus doubly glycosylated protein[5]. This in turn can be converted into an apparent free energy of membrane insertion, ΔG_{app} (see Methods). Control experiments show that the identity of the TM1 and TM2 helices in Lep has little influence on ΔG_{app} (ref. 8).

Position-dependent contributions to ΔG_{app}

To obtain a comprehensive data set describing the positional variability in ΔG^{aa}_{app} for the 20 amino acids, we designed a set of Lep constructs in which each kind of residue was systematically scanned across a Leu-Ala-based H-segment and ΔG_{app} values were measured. All H-segments were designed with the sequence GGPG-(19 residues)-GPGG. For each residue type, the numbers of Leu and Ala residues in the H-segment were chosen such that $\Delta G_{app} \approx 0$ kcal mol^{-1} (1 kcal ≈ 4.18 kJ) when the residue was in the middle position of the 19-residue stretch. The results show that ΔG^{aa}_{app} values vary strongly with position for charged and highly polar

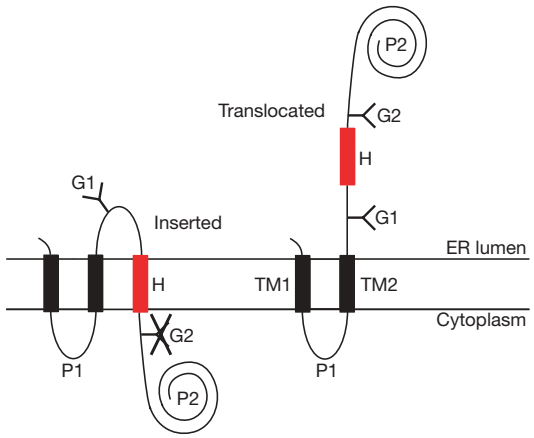


Figure 1 | The Lep model protein. *Escherichia coli* leader peptidase (Lep) has two TM helices (TM1 and TM2) and a large luminal domain (P2). It inserts into rough microsomes in an N_{lum}–C_{lum} orientation. H-segments (red) are engineered into the P2 domain with two flanking Asn-X-Thr glycosylation acceptor sites (G1, G2). Constructs for which the H-segment is integrated into the endoplasmic reticulum membrane as a TM helix are glycosylated only on the G1 site (left), whereas those for which the H-segment is translocated across the membrane are glycosylated on both the G1 and G2 sites (right).

[1]Center for Biomembrane Research, Department of Biochemistry and Biophysics, Stockholm University, SE-106 91 Stockholm, Sweden. [2]Stockholm Bioinformatics Center, AlbaNova, Stockholm University, SE-106 91 Stockholm, Sweden. [3]Department of Physiology and Biophysics and the Center for Biomembrane Systems, University of California at Irvine, Irvine, California 92697-4560, USA.
*These authors contributed equally to this work.

residues as well as for Pro (a strong helix-breaker), whereas they are nearly independent of position for weakly polar and apolar residues (Supplementary Fig. 1a). Although single charged residues may cause a shift in the position of the H-segment relative to the membrane and lead to an underestimate of ΔG^{aa}_{app}, previous work suggests that such shifts are small (not more than 3 Å) and restricted to positions near the ends of the 19-residue H-segments used here[9,10].

Although the single-residue scans give a good first impression of the positional dependence of the ΔG^{aa}_{app} values, we sought to incorporate as much information as possible, both from the H-segments analysed previously[5,7] and from the constructs made in this study (Supplementary Table 1), to derive an optimized matrix ($\overline{\Delta G^{aa}_{app}}$) of position-specific ΔG^{aa}_{app} values. To this end, 324 19-residue H-segments for which the measured ΔG_{app} values are between -1.5 and $+1.0$ kcal mol^{-1} (that is, within an interval where the accuracy in the ΔG_{app} determination is good) were collected. Using this data set, we performed a least-squares optimization in which the $\overline{\Delta G^{aa}_{app}}$ matrix elements for each residue were described by a gaussian function (see Methods). We also included a contribution from the hydrophobic moment of each H-segment[5]. Except for the hydrophobic moment part, H-segment ΔG_{app} values were modelled as a linear sum of free-energy values for individual amino acids:

$$\Delta G^{pred}_{app} = \sum_{i=1}^{l} \Delta G^{aa(i)}_{app} + c_0\mu \qquad (1)$$

where l is the length of the segment (here, $l = 19$), $\Delta G^{aa(i)}_{app}$ is the matrix element giving the contribution from amino acid aa in position i, μ is the hydrophobic moment (see Methods), and c_0 is the weight parameter for the hydrophobic moment. The optimized $\overline{\Delta G^{aa}_{app}}$ matrix (Supplementary Table 2) was derived by minimizing the sum of the squared differences between the predicted ΔG_{app} values (ΔG^{pred}_{app}) and measured ΔG_{app} values.

As expected, equation (1) reproduces the experimental single-residue scans well (Supplementary Fig. 1a) and also reproduces data from symmetrical pair-scans[5] in which two residues were scanned symmetrically from the centre of the H-segment to preclude shifts in the location of the H-segment relative to the membrane (Supplementary Fig. 1b). There is only one residue, proline, for which the single-scan and pair-scan results are qualitatively different: Pro has a fairly symmetric single-scan profile, but two Pro residues placed near each other in the centre of the H-segment are tolerated better than when they are spaced farther apart. This cooperative effect cannot be captured by the simple additive model in equation (1).

The optimized position-dependent gaussians describing the $\overline{\Delta G^{aa}_{app}}$ matrix are shown in Fig. 2 (blue curves). Figure 2 also shows statistical free-energy profiles derived from the distribution of the different amino acids in high-resolution membrane protein three-dimensional structures (red curves; see Methods); these profiles presumably reflect mainly interaction free energies between amino-acid side chains and the lipid bilayer. In general, the two sets of profiles match each other well. The profiles for His match rather poorly, however. A possible explanation is that a number of the known three-dimensional structures contain cofactor-binding His residues. Indeed, the statistical His profile obtained when all such

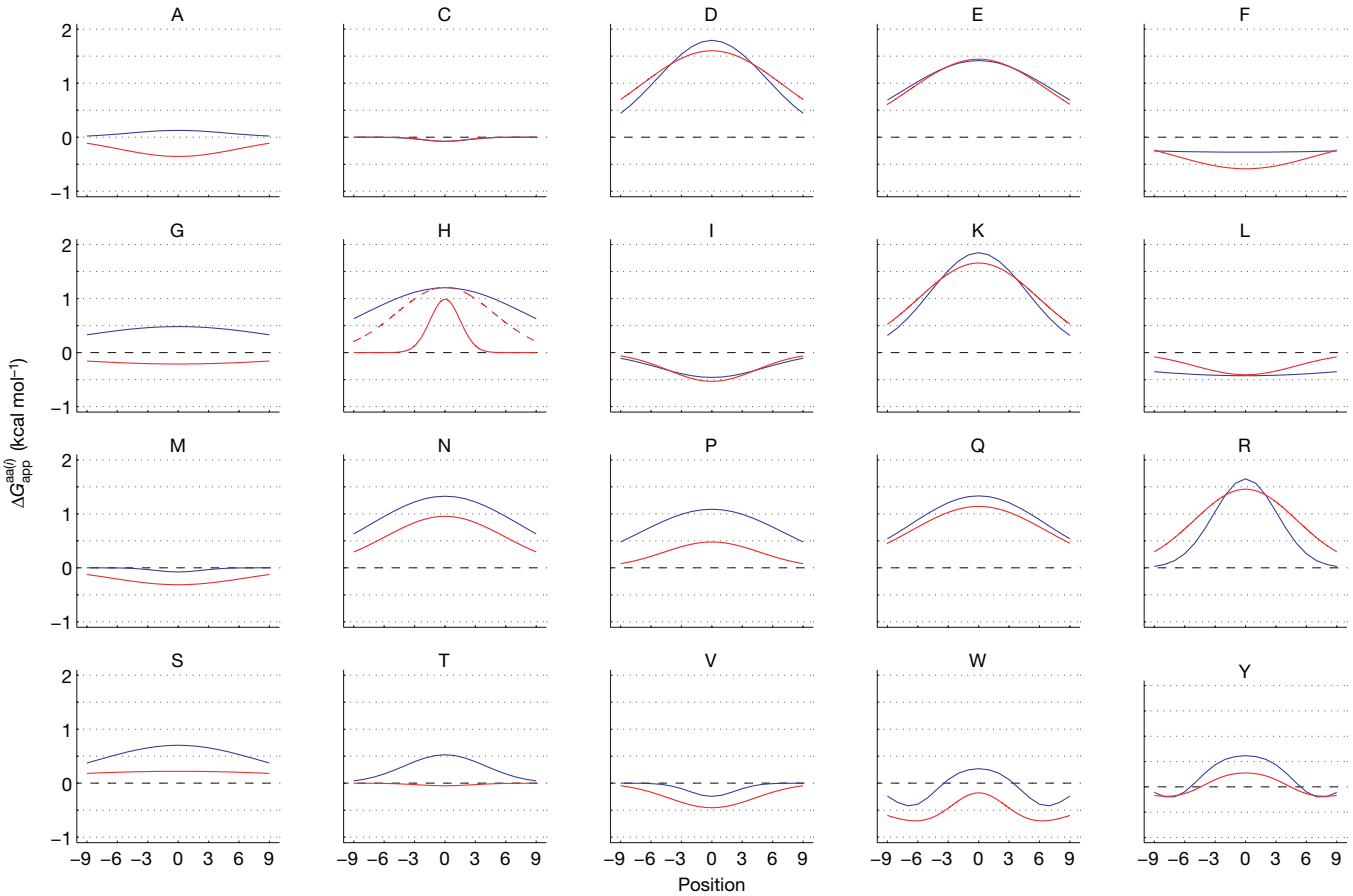


Figure 2 | Position-specific ΔG_{app} contributions. The gaussians describing the $\overline{\Delta G^{aa}_{app}}$ matrix—that is, the contribution from each individual amino acid to ΔG_{app}—are plotted as a function of position within the 19-residue segment (blue). Amino acids are identified by their one-letter abbreviations. Position-specific statistical distributions calculated from three-dimensional structures of membrane proteins are shown in red. The dashed red line for His shows the statistical distribution obtained when all cofactor-containing proteins are omitted. To compare the profiles, one amino acid was equated to a z-coordinate displacement of 1.5 Å.

proteins are omitted matches the experimental profile much better (dashed red line).

How well can equation (1) predict ΔG_{app} values for H-segments in the training set or chosen from natural proteins? For 90% of the H-segments in the training set, ΔG_{app}^{pred} values are within ± 0.45 kcal mol^{-1} of the measured ΔG_{app} values (Supplementary Fig. 2). Overall, ΔG_{app} is well predicted by equation (1) also for 16 representative 19-residue segments from membrane proteins of known structure (see below) and six additional segments that include sequences from a weakly hydrophobic single-span TM protein and five non-membrane proteins (Supplementary Fig. 2). The only outliers are a couple of very highly charged S4 helices from ion-channel voltage-sensor domains[11], for which equation (1) overestimates the cost of membrane insertion. We conclude that the simple additive model provides a good first approximation to ΔG_{app} for most natural protein sequences, unless they are exceptionally rich in charged residues or contain multiple proline residues.

Relation between H-segment length and ΔG_{app}

To delineate the relation between H-segment length and ΔG_{app}, we analysed a series of Leu-Ala based constructs with the overall composition GGPG-(*n*L, *m*A)-GPGG, where $n = 0$, 1, 2, 3, 5, 7. The *m* values were chosen such that we could identify by interpolation, for each *n*, the *m* value for which the H-segment inserts into the membrane in 50% of the molecules (m_{50}, corresponding to $\Delta G_{app} = 0$ kcal mol^{-1}). In addition, we included one set of constructs with the overall composition GGPG-(*n*L)-GPGG. The results are shown in Fig. 3. As the number of Leu residues (*n*) decreases, the number of Ala residues (*m*) required to reach a given ΔG_{app} increases; in fact, as shown in the inset to Fig. 3, there is a striking linear correlation between the number of Leu and Ala residues in the H-segment required for $\Delta G_{app} = 0$ kcal mol^{-1}. The least-squares fit to the data points in Fig. 3 (inset) is $m_{50} = -2.9n + 26$; that is, for each Leu residue removed from the H-segment, 2.9 Ala have to be added to maintain $\Delta G_{app} = 0$ kcal mol^{-1}. The correlation holds over an extended range of H-segment lengths ($9 \leq n + m \leq 30$). Because the Leu side chain has a roughly 2.4-fold larger accessible surface area than the Ala side chain[12] (95 Å^2 compared with 40 Å^2), the Leu-Ala based H-segments with $\Delta G_{app} = 0$ kcal mol^{-1} have a roughly constant non-polar surface area of about 1,000 Å^2.

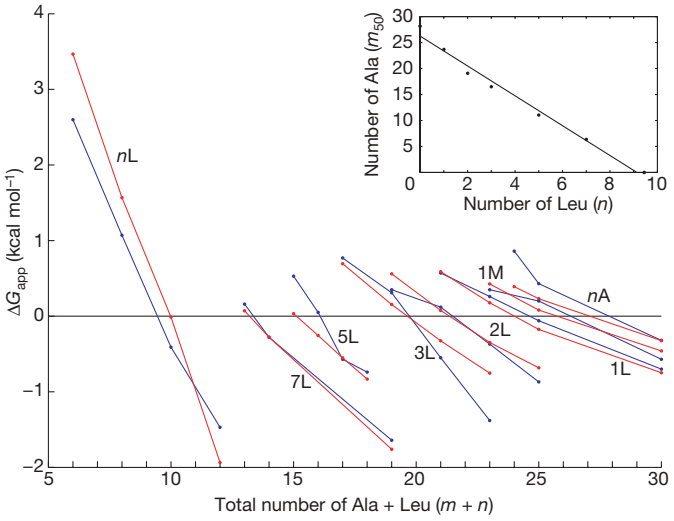


Figure 3 | Length dependence of ΔG_{app}. Measured ΔG_{app} (blue) and predicted, cross-validated ΔG_{app}^{pred} values (red) for H-segments with the composition GGPG-(*n*L, *m*A)-GPGG and GGPG-(1M, *m*A)-GPGG. The lines connect data points with fixed *n* and varying *m*. The inset plots the number of Ala residues against the number of Leu residues required for $\Delta G_{app} = 0$ kcal mol^{-1} obtained from the data in the main panel ($m_{50} = -2.9n + 26$; $R^2 = 0.99$). The data for the (1M, *n*A) constructs were not used in the optimization of the length-corrected equation (1).

A second notable feature in Fig. 3 is that the slope of the lines for different *n* tends towards zero as the overall length $l = n + m$ increases. Closer inspection reveals that the derivative $\partial \Delta G_{app}/\partial l$ is roughly proportional to l (Supplementary Fig. 3). This suggests that a phenomenological, length-dependent expression for ΔG_{app}^{pred} can be obtained from equation (1) (which holds for $l = 19$) to which is added an expression of the form $c_1 + c_2 l + c_3 l^2$; the best fit obtained by optimizing c_1, c_2 and c_3 (see Methods) is shown as red lines in Fig. 3.

To map the variation in $\Delta G_{app}^{aa(i)}$ values as a function of H-segment length, we measured position-dependent ΔG_{app} values for a single Lys residue in H-segments of different lengths (15, 19 and 25 residues; Supplementary Fig. 4a). A lengthening of the H-segment essentially results in a 'stretching' of the Lys profile, while maintaining the difference in ΔG_{app} between the middle and terminal positions at about 1.8 kcal mol^{-1}. We further tested whether the $\Delta G_{app}^{aa(i)}$ values for residues spaced at either end of a long H-segment are additive or whether there is a maximum length over which two residues cannot simultaneously contribute to ΔG_{app}. To this end, we scanned two Leu residues symmetrically from the centre of a 25-residue H-segment. As found for a 19-residue H-segment[5], ΔG_{app} is roughly constant regardless of the spacing between the two Leu residues (Supplementary Fig. 4b). We also found that the contributions to ΔG_{app} from Trp residues introduced near the ends of the H-segment are roughly additive for both 19-residue and 25-residue H-segments (constructs 40, 363 and 417–422 in Supplementary Table 1). These results imply that even very long H-segments behave as one unit in terms of membrane insertion and that equation (1), corrected for length dependence and with the $\overline{\Delta G_{app}^{aa}}$ matrix suitably 'stretched' or 'compressed' for different lengths (see Methods), should provide a good model for TM helix recognition by the Sec61 translocon. A web server implementing the length-corrected model for ΔG_{app}^{pred} is available at http://www.cbr.su.se/DGpred/.

Contributions to ΔG_{app} from flanking residues

The cytoplasmic ends of TM helices are often flanked by positively charged Lys and Arg residues[13], suggesting that flanking charged residues might contribute to ΔG_{app}. We therefore tested two series of H-segments in which the central 19-residue stretch had the composition 3L/16A or 4L/15A. The Gly residues in the GGPG…GPGG flanks used above were replaced with Asp, Glu, Asn, Gln, Lys, Arg or Ser, and the luminal and cytoplasmic flanks were combined in different ways.

The changes in ΔG_{app} relative to the 3L/16A and 4L/15A H-segments with GGPG…GPGG flanks are shown in Supplementary Fig. 5a. The effects of Asp and Glu, as well as those of Asn and Gln, are strikingly different from the effects of Lys and Arg. Three Asp/Glu or Asn/Gln residues increase ΔG_{app} by about 0.9 kcal mol^{-1} and about 0.5 kcal mol^{-1}, respectively, when present at the luminal end of the H-segment but not at its cytoplasmic end. In contrast, three Lys or Arg residues reduce ΔG_{app}, both by -0.7 kcal mol^{-1}, when present at the cytoplasmic end but not at the luminal end. Ser has no appreciable effect on ΔG_{app} compared with Gly. The contributions to ΔG_{app} from flanking Asp and Lys residues are approximately additive, such that $\Delta\Delta G_{app}$ for H-segments with DDPD…KPKK flanks is close to that expected from adding the individual contributions (expected average $\Delta\Delta G_{app} = (0.8 - 0.7) = 0.1$ kcal mol^{-1}; observed average $\Delta\Delta G_{app} = 0.2$ kcal mol^{-1}). Additivity implies that the entire H-segment, including the flanking residues, may be recognized as one unit during membrane insertion.

To check whether this conclusion is valid also for other H-segment lengths, we made constructs with the same combinations of flanking residues for a 10-residue (10L) and a 25-residue (2L/23A) H-segment. As is clear from Supplementary Fig. 5b, the contributions from the charged flanking residues are additive even for the 25-residue

H-segment. Thus, flanking residues separated by as few as 10 and as many as 25 apolar residues (that is, spaced 15–38 Å apart) affect Sec61-mediated membrane integration of TM helices in the same way.

ΔG^{pred}_{app} for TM helices in natural proteins

Finally, how well does the length-corrected equation (1) serve to identify TM helices in natural proteins? To address this question, we collected four test sets: first, all mammalian proteins annotated in SwissProt[14] as having a cleaved signal peptide and one single TM helix (349 single-spanning membrane proteins); second, all mammalian proteins annotated as having a cleaved signal peptide and no TM helix (1,012 soluble proteins targeted to the secretory pathway); third, all mammalian proteins annotated as located in the cytoplasm (670 cytoplasmic proteins); and fourth, all helix-bundle membrane proteins of known three-dimensional structure with at least two TM helices (508 TM helices from 66 Protein Data Bank[15] structures). The first, second and fourth sets contain proteins that have all passed through the endoplasmic reticulum translocon (or its prokaryotic SecYEG homologue), whereas the proteins in the third set have not visited the translocon. For the first to third sets, we identified in each protein (after removing the signal peptide) the segment with the lowest ΔG^{pred}_{app} value (for $17 \leq l \leq 33$), whereas for the fourth set we identified the segment with the lowest ΔG^{pred}_{app} value within each annotated TM helix (extended by ten residues at both the amino-terminal end and the carboxy-terminal end).

The results are summarized in Fig. 4. The overlap between the ΔG^{pred}_{app} distributions for the single-spanning transmembrane proteins and the secreted proteins is small, and the two distributions cross close to the zero-point on the scale defined by the experimental analysis of the designed H-segments. The discrimination between the two data sets is considerably better with the use of the ΔG^{pred}_{app} values than when simpler hydrophobicity scales are used (Supplementary Fig. 6).

The ΔG^{pred}_{app} distribution for the cytoplasmic proteins overlaps for the most part with that for the secreted proteins, as expected. There is, however, a significant number of cytoplasmic proteins with $\Delta G^{pred}_{app} < 0\,\mathrm{kcal\,mol^{-1}}$, as though the requirement to pass through the translocon has 'filtered out' proteins with such segments from the group of secreted proteins.

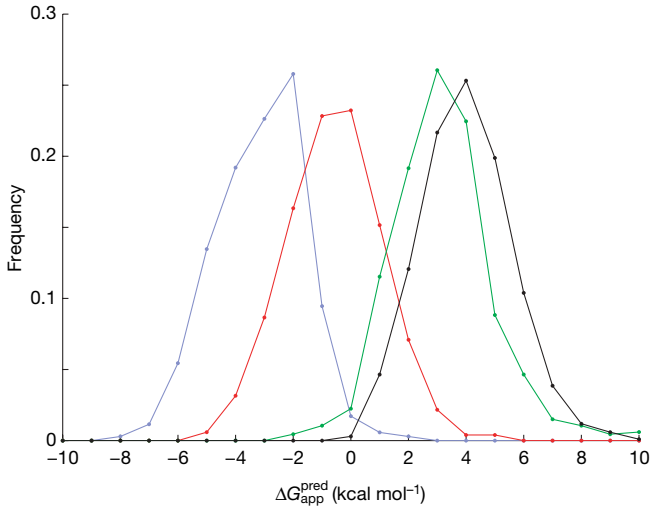


Figure 4 | Distributions of ΔG^{pred}_{app} values in natural proteins. The segment with lowest ΔG^{pred}_{app} was identified in 670 cytoplasmic (green), 1,012 secreted (black) and 349 single-spanning transmembrane proteins (blue); signal peptides were excluded. For 508 TM helices from multispanning membrane proteins of known three-dimensional structure (red), the segment with the lowest ΔG^{pred}_{app} for each helix was identified. Dots show the relative frequency of proteins with ΔG^{pred}_{app} within $\pm 0.5\,\mathrm{kcal\,mol^{-1}}$ of the value given on the *x* axis. None of the 1,012 secreted proteins has a segment with $\Delta G^{pred}_{app} < 0$, and only 3 of 349 transmembrane segments in the single-spanning proteins have $\Delta G^{pred}_{app} > 0$.

Although more data will be required for proper modelling of the quantitative effects on ΔG^{pred}_{app} of charged flanking residues, a rough estimate is that, on average, cytoplasmic positively charged flanking residues may decrease ΔG^{pred}_{app} by about $0.5\,\mathrm{kcal\,mol^{-1}}$ (M.L.-B., C. Lundin, H.K., I.N. and G.v.H., unpublished observations). Even with this correction, however, there is a surprisingly large fraction (about 25%) of the TM helices in the multi-spanning membrane proteins of known three-dimensional structure that have $\Delta G^{pred}_{app} > 0\,\mathrm{kcal\,mol^{-1}}$. Such segments would presumably be only inefficiently recognized as TM helices by the translocon if they were the only hydrophobic segment in a protein (as seen for the few that were tested in Supplementary Fig. 2). This suggests that a relatively large fraction of the TM helices in multi-spanning membrane proteins may depend on interactions with neighbouring TM helices for proper partitioning into the membrane. Indeed, several such cases have been described in the literature[16,17].

Our results show that the experimentally derived position-dependent $\Delta G^{aa(i)}_{app}$ profiles are similar to statistical residue-distribution profiles derived from TM helices in natural membrane proteins of known structure. ΔG^{pred}_{app} values obtained from a simple additive model, equation (1), are reasonably close to the ΔG_{app} values measured for H-segments of mixed amino-acid composition extracted from natural proteins, and they provide a good discrimination between TM helices in single-spanning mammalian membrane proteins and the most hydrophobic segments in mammalian secreted proteins. The relation between length and hydrophobicity for membrane insertion of Ala/Leu-based H-segments is a strikingly simple one, and H-segments as long as 25 residues behave as a single unit during membrane insertion; the simplest interpretation is that long H-segments can tilt or flex and thereby interact in their entirety with the lipid bilayer despite considerable 'hydrophobic mismatch'[18–21] and that the length-dependent contribution to ΔG_{app} approximates the free-energy cost associated with positive and negative mismatch between helix length and bilayer thickness. These results further support the idea that the recognition of TM helices by the Sec61 translocon critically involves a partitioning of the nascent polypeptide into the lipid bilayer[5,22], and they provide a quantitative basis for future studies of membrane protein biogenesis and prediction of membrane protein topology and structure.

METHODS SUMMARY

Lep constructs were transcribed and translated in the TNT Quick coupled transcription–translation system supplemented with dog pancreas rough microsomes. The degree of membrane integration of each H-segment was quantified from SDS–PAGE gels by calculating an apparent equilibrium constant between the membrane-integrated and non-integrated forms: $K_{app} = f_{1g}/f_{2g}$, where f_{1g} is the fraction of singly glycosylated Lep molecules and f_{2g} is the fraction of doubly glycosylated Lep molecules, after correcting for the fact that a fully translocated P2 domain is glycosylated only to about 85% (ref. 5). The results were then converted to apparent free energies, $\Delta G_{app} = -RT\ln K_{app}$.

The full expression for the length-corrected equation (1) is

$$\Delta G^{pred}_{app} = \sum_{i=1}^{l} \Delta G^{aa(i)}_{app} + c_0 \sqrt{\left(\sum_{i=1}^{l} \Delta G^{aa(i)}_{app} \sin(100^\circ i)\right)^2 + \left(\sum_{i=1}^{l} \Delta G^{aa(i)}_{app} \cos(100^\circ i)\right)^2} + c_1 + c_2 l + c_3 l^2$$

with the optimized parameter values given in Supplementary Table 2.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 9 August; accepted 17 October 2007.

1. Oberai, A., Ihm, Y., Kim, S. & Bowie, J. U. A limited universe of membrane protein families and folds. *Protein Sci.* **15**, 1723–1734 (2006).
2. Wiener, M. C. & White, S. H. Structure of a fluid dioleoylphosphatidylcholine bilayer determined by joint refinement of x-ray and neutron diffraction data. III. Complete structure. *Biophys. J.* **61**, 437–447 (1992).
3. Ulmschneider, M. B., Sansom, M. S. & Di Nola, A. Properties of integral membrane protein structures: derivation of an implicit membrane potential. *Proteins* **59**, 252–265 (2005).

4. Schnell, D. J. & Hebert, D. N. Protein translocons: multifunctional mediators of protein translocation across membranes. *Cell* **112,** 491–505 (2003).
5. Hessa, T. *et al.* Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* **433,** 377–381 (2005).
6. Heinrich, S., Mothes, W., Brunner, J. & Rapoport, T. The Sec61p complex mediates the integration of a membrane protein by allowing lipid partitioning of the transmembrane domain. *Cell* **102,** 233–244 (2000).
7. Hessa, T., White, S. H. & von Heijne, G. Membrane insertion of a potassium channel voltage sensor. *Science* **307,** 1427 (2005).
8. Meindl-Beinker, N. M., Lundin, C., Nilsson, I., White, S. H. & von Heijne, G. Asn- and Asp-mediated interactions between transmembrane helices during translocon-mediated membrane protein assembly. *EMBO Rep.* **7,** 1111–1116 (2006).
9. Nilsson, I. *et al.* Proline-induced disruption of a transmembrane α-helix in its natural environment. *J. Mol. Biol.* **284,** 1165–1175 (1998).
10. Monné, M., Nilsson, I., Johansson, M., Elmhed, N. & von Heijne, G. Positively and negatively charged residues have different effects on the position in the membrane of a model transmembrane helix. *J. Mol. Biol.* **284,** 1177–1183 (1998).
11. Zhang, L. *et al.* Membrane insertion of the Shaker voltage sensor occurs both cotranslationally and posttranslationally. *Proc. Natl Acad. Sci. USA* **104,** 8263–8268 (2007).
12. Chothia, C. The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* **105,** 1–12 (1976).
13. von Heijne, G. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. *EMBO J.* **5,** 3021–3027 (1986).
14. O'Donovan, C. *et al.* High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief. Bioinform.* **3,** 275–284 (2002).
15. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).
16. Sadlish, H. & Skach, W. R. Biogenesis of CFTR and other polytopic membrane proteins: new roles for the ribosome–translocon complex. *J. Membr. Biol.* **202,** 115–126 (2004).
17. Buck, T. M., Wagner, J., Grund, S. & Skach, W. R. A novel tripartite motif involved in aquaporin topogenesis, monomer folding and tetramerization. *Nature Struct. Mol. Biol.* **14,** 762–769 (2007).
18. Killian, J. A. & von Heijne, G. How proteins adapt to a membrane–water interface. *Trends Biochem. Sci.* **25,** 429–434 (2000).
19. de Planque, M. R. R. & Killian, J. A. Protein–lipid interactions studied with designed transmembrane peptides: role of hydrophobic matching and interfacial anchoring. *Mol. Membr. Biol.* **20,** 271–284 (2003).
20. Yeagle, P. L., Bennett, M., Lemaitre, V. & Watts, A. Transmembrane helices of membrane proteins may flex to satisfy hydrophobic mismatch. *Biochim. Biophys. Acta* **1768,** 530–537 (2007).
21. Monné, M. & von Heijne, G. Effects of 'hydrophobic mismatch' on the location of transmembrane helices in the ER membrane. *FEBS Lett.* **496,** 96–100 (2001).
22. Heinrich, S. U. & Rapoport, T. A. Cooperation of transmembrane segments during the integration of a double-spanning protein into the ER membrane. *EMBO J.* **22,** 3654–3663 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank E. Missioux for technical assistance, and A. Elofsson and E. Lindahl for discussions. This work was supported by grants from the Swedish Foundation for Strategic Research, the Marianne and Marcus Wallenberg Foundation, the Swedish Cancer Foundation, the Swedish Research Council and the European Commission (BioSapiens) to G.v.H., the Magnus Bergvall Foundation to I.N., the National Institute of General Medical Sciences to S.H.W., the Swiss National Science Foundation to M.L.-B., and the Japan Society for the Promotion of Science to Y.S.

Author Contributions T.H. and N.M.M.-B. performed the experimental work together with H.K., Y.S., M.L.-B. and I.N. A.B. performed the computational work. T.H., N.M.M.-B., A.B., S.H.W. and G.v.H. prepared the manuscript. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to G.v.H. (gunnar@dbb.su.se).

METHODS

Enzymes and chemicals. All enzymes, plasmid pGEM1, and the TNT Quick transcription–translation system were from New England Biolabs or Promega. [^{35}S]Met, deoxynucleotides and dideoxyribonucleotides were from GE Healthcare. Oligonucleotides were from Cybergene and MWG Biotech.

Expression *in vitro* and quantification of membrane insertion efficiency. All plasmids were constructed as described[23]. Constructs cloned in pGEM1 were transcribed and translated in the TNT Quick coupled transcription–translation system. The reaction was started by the addition of 1 μg of DNA template, 1 μl of [^{35}S]Met (15 μCi), and 1 μl of dog pancreas rough microsomes (a gift from M. Sakaguchi), and samples were incubated for 90 min at 30 °C. Translation products were analysed by SDS–PAGE and gels were quantified on a Fuji FLA-3000 PhosphorImager with the use of Image Reader 8.1j software. The degree of membrane integration of each H-segment was quantified from SDS–PAGE gels by calculating an apparent equilibrium constant between the membrane-integrated and non-integrated forms: $K_{app} = f_{1g}/f_{2g}$, where f_{1g} is the fraction of singly glycosylated Lep molecules and f_{2g} is the fraction of doubly glycosylated Lep molecules after correcting for the fact that a fully translocated P2 domain is only about 85% glycosylated[23]. The results were then converted to apparent free energies, $\Delta G_{app} = -RT \ln K_{app}$. All ΔG_{app} values were calculated as mean values from at least two independent experiments.

Optimization of position-specific ΔG_{app} contributions. Position-specific residue contributions to ΔG_{app} were calculated by using an additive model with an additional term to account for the hydrophobic moment[24] (μ) of the H-segment:

$$\Delta G_{app}^{pred} = \sum_{i=1}^{l} \Delta G_{app}^{aa(i)} + c_0 \sqrt{\left(\sum_{i=1}^{l} \Delta G_{app}^{aa(i)} \sin(100^\circ i)\right)^2 + \left(\sum_{i=1}^{l} \Delta G_{app}^{aa(i)} \cos(100^\circ i)\right)^2} \quad (2)$$

All amino acid profiles except those for Trp and Tyr were appoximated by single gaussians with two parameters:

$$\Delta G_{app}^{aa(i)} = a_0^{aa} e^{-a_1^{aa} i^2} \quad (3)$$

where i denotes the position in the H-segment, with $i = 0$ corresponding to the central residue. To reproduce the characteristic 'W' shape of the single-scan curves for Trp and Tyr (see Supplementary Fig. 1), double gaussians (five parameters) were used for these two profiles:

$$\Delta G_{app}^{aa(i)} = a_0^{aa} e^{-a_1^{aa} i^2} + a_2^{aa}(e^{-a_3^{aa}(i-a_4^{aa})^2} + e^{-a_3^{aa}(i+a_4^{aa})^2}) \quad (4)$$

Finally, the sum of squares of the differences between measured and predicted ΔG_{app} values was minimized:

$$\hat{\Theta} = \arg \min_{\Theta} \left(\sum_{\text{All constructs}} (\Delta G_{app}^{pred} - \Delta G_{app})^2 \right) \quad (5)$$

where the sum goes over all constructs, ΔG_{app} is the experimentally measured value, ΔG_{app}^{pred} is the predicted value according to equation (2), and $\Theta = \{a_0^{aa}, a_1^{aa} \ldots a_4^{aa}, c_0\}$ is the set of 47 parameters (46 a parameters needed to describe the 20 profiles according to equations (3) and (4), plus the additional hydrophobic moment weight parameter c_0). An underlying assumption using the above parameterization is that the profiles are symmetric around the middle of the membrane. Judging from the experimental profiles (Supplementary Fig. 1), this assumption seems justified.

A total of 321 19-residue H-segments with ΔG_{app} values between −1.5 and +1.0 kcal mol^{-1} plus 3 H-segments from the Arg scan with ΔG_{app} values slightly lower than −1.5 kcal mol^{-1} were used in the optimization (see Supplementary Table 1). The interval is asymmetric because the degree of double glycosylation of fully translocated H-segments can vary slightly between different batches of microsomes, introducing an extra variability in the measurements of high ΔG_{app} values.

As the starting point for the optimization, all $\Delta G_{app}^{aa(i)}$ values were set equal to zero. In a first pre-optimization step, the position-specific $\Delta G_{app}^{aa(i)}$ values were treated as being independent; that is, without the parameterization as in equations (3) and (4). The optimization was thus performed with respect to the full (20 × 19) $\overline{\Delta G_{app}^{aa}}$ matrix. Gaussian functions were then fitted to the resulting curves, and the corresponding parameter values were used as the starting point for the final optimization of the 47 parameters in Θ, now using the parameterization as given by equations (3) and (4). The pre-optimization step thus served to quickly find the approximate region in parameter space for the final solution.

To estimate the ability of the model to predict ΔG_{app} values of constructs outside the training set, we performed a leave-one-out cross-validation procedure in which the model was trained on all constructs except one and then used to predict the ΔG_{app} value of the missing construct from equation (2) (Supplementary Fig. 2).

For the optimization, we used the MATLAB v. 7.0.1 (MathWorks Inc.) implementation of a subspace trust region method[25] that iteratively searches for a local minimum to the minimization criterion equation (5) based on gradient descent. The optimized $\overline{\Delta G_{app}^{aa}}$ matrix as well as the optimized Θ and length parameter values are given in Supplementary Table 2, and the resulting profiles are shown in Fig. 2.

We also tried another, previously used function[26] to fit the experimental residue-distribution profiles. The resulting profiles are similar to the gaussian profiles, and the correlation between the predicted and experimental ΔG_{app} values is essentially the same as with the gaussian profiles ($R^2 = 0.78$ versus 0.79). Because the function used in ref. 26 requires 67 parameters to describe the residue-distribution profiles (against 47 for the gaussian functions) we kept the simpler gaussian representation.

Influence of H-segment length on ΔG_{app}. Because the derivative of ΔG_{app} with respect to segment length, $\partial \Delta G_{app}/\partial l$, increases roughly in proportion to l (Supplementary Fig. 3), we assumed that a general quadratic expression accounts for the length dependence of ΔG_{app}. On the basis of measured ΔG_{app} values for all 19-residue constructs analysed above but now also including the ones of variable length with $\Delta G_{app} \in [-1,1]$ kcal mol^{-1} (Fig. 3), we minimized the following criterion by using the same optimization algorithm as above:

$$[\hat{c}_1, \hat{c}_2, \hat{c}_3] = \arg \min_{c_1,c_2,c_3} \left(\sum_{\text{All constructs}} (\Delta G_{app}^{pred} - \Delta G_{app} + c_1 + c_2 l + c_3 l^2)^2 \right) \quad (6)$$

where ΔG_{app}^{pred} is the predicted value according to equation (1) (that is, without consideration of length), ΔG_{app} is the measured value and c_1, c_2, and c_3 are parameters describing the length dependence (their optimized values are given in Supplementary Table 2). The final model used for predicting natural segments therefore contained a total of 50 parameters (47 from equation (5) plus three length parameters).

To obtain ΔG_{app}^{pred} for segments with $l \neq 19$, 'stretched' or 'compressed' $\Delta G_{app}^{aa(i)}$ profiles were used. Because the original $\Delta G_{app}^{aa(i)}$ profile values were calculated for $l = 19$, the position coordinate $j = 1, \ldots, k$ of a segment of length $k \neq 19$ was first transformed into the native coordinate system $i = -9, \ldots, +9$ (used in, for example, Fig. 2) using the expression $i = 9\{2[(j-1)/(k-1)] - 1\}$, and then the original profiles for $l = 19$ were applied.

The predicted values in Fig. 3 and Supplementary Table 1 were obtained with the leave-one-out cross-validation procedure.

Statistical distributions of amino acids in natural transmembrane helices. To compare the optimized $\Delta G_{app}^{aa(i)}$ profiles with statistical distributions from natural TM helices, we calculated residue distributions along the membrane normal for 575 TM helices in 158 non-redundant chains from 77 high-resolution X-ray structures[27]. Homology reduction at an 80% sequence identity threshold was performed with the CD-HIT algorithm[28]. Residue frequencies along the membrane normal were calculated and divided into bins 1.5 Å wide with respect to the distance from the membrane centre. Statistical $\Delta G_{stat}^{aa(i)}$ profiles for residue distributions in the [−45, +45] Å distance interval were then calculated as

$$\Delta G_{stat}^{aa(i)} = -RT \ln \left(\frac{f(aa,i)}{bgr(aa)} \right) \quad (7)$$

where f(aa,i) is the frequency of amino acid aa in helix position i normalized such that the sum over all amino acids adds to 1, and bgr(aa) is the background frequency of amino acid aa, according to the amino acid composition of SwissProt[29] (version 50.1), resembling the procedures used in similar studies[26,30]. Finally, to smooth the profiles, least-squares curve fitting was performed in accordance with equations (3) and (4). To compare the profiles with the $\overline{\Delta G_{app}^{aa}}$ matrix it was assumed that each amino acid in the H-segments corresponded to a z-coordinate displacement of 1.5 Å. Although the curve fitting was performed with respect to the full [−45, +45] Å interval, in Fig. 2 only the [−13.5, +13.5] Å interval of the curves is shown.

Distributions of ΔG_{app}^{pred} values in natural proteins. To investigate how the distributions of ΔG_{app}^{pred} values differ between sets of secreted, transmembrane and cytoplasmic proteins, we compiled four data sets that were all homology-reduced at an 80% sequence identity threshold with the CD-HIT algorithm[28]: first, all mammalian proteins in SwissProt[29] (version 50.1) annotated to have a signal peptide and exactly one transmembrane region (349 single-spanning membrane proteins); second, all mammalian proteins in SwissProt annotated to have a signal peptide but no transmembrane regions (1,012 soluble proteins targeted to the secretory pathway); third, all mammalian proteins in SwissProt annotated with 'cytoplasm' as subcellular location (670 cytoplasmic proteins); and fourth, all known X-ray structures of membrane proteins from the OPM database[27] with at least two TM helices (66 multi-spanning membrane proteins with a total of 508 TM helices). Proteins annotated in SwissProt as having a GPI anchor were removed before the analysis, because they were in

some cases annotated as having a transmembrane segment and in some cases not. Annotated signal peptides were removed from all sequences.

A sliding-window approach was employed to identify the segment of length 17–33 residues with the lowest ΔG_{app}^{pred}. In the first to third sets we scanned the full protein sequences, whereas in the fourth set we extended each annotated TM helix by ten residues on both the amino-terminal end and the carboxy-terminal end and then scanned for such a segment.

To compare the ΔG_{app}^{pred} predictions against existing hydrophobicity-based predictions, a similar sliding-window analysis, but with a fixed window length ($l = 19$), was also performed with the Zhao–London[31], Kyte–Doolittle[32] and Wimley–White[33] hydrophobicity scales. Because these scales do not contain position-dependent information, for comparison we also made predictions with a simpler version of the 'biological' scale, in which all profiles were replaced with their respective mean $\Delta G_{app}^{aa(i)}$ value (that is, no positional dependence), and in addition the terms modelling length and hydrophobic moment were left out (c_0, c_1, c_2 and c_3 were set to zero, and the window length was fixed at $l = 19$). The resulting distributions are shown in Supplementary Fig. 6. If l is allowed to vary between 17 and 33 residues (as in the ΔG_{app}^{pred} calculations), the overlaps between the distributions for single-spanning and multi-spanning membrane change only slightly (data not shown).

23. Hessa, T. *et al.* Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* **433,** 377–381 (2005).
24. Eisenberg, D., Schwarz, E., Komaromy, M. & Wall, R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.* **179,** 125–142 (1984).
25. Coleman, T. F. & Li, Y. An interior, trust region approach for nonlinear minimization subject to bounds. *SIAM J. Optimiz.* **6,** 418–445 (1996).
26. Senes, A. *et al.* E_z, a depth-dependent potential for assessing the energies of insertion of amino acid side-chains into membranes: Derivation and applications to determining the orientation of transmembrane and interfacial helices. *J. Mol. Biol.* **366,** 436–448 (2007).
27. Lomize, M. A., Lomize, A. L., Pogozheva, I. D. & Mosberg, H. I. OPM: orientations of proteins in membranes database. *Bioinformatics* **22,** 623–625 (2006).
28. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22,** 1658–1659 (2006).
29. O'Donovan, C. *et al.* High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief. Bioinform.* **3,** 275–284 (2002).
30. Ulmschneider, M. B., Sansom, M. S. & Di Nola, A. Properties of integral membrane protein structures: derivation of an implicit membrane potential. *Proteins* **59,** 252–265 (2005).
31. Zhao, G. & London, E. An amino acid 'transmembrane tendency' scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: Relationship to biological hydrophobicity. *Prot. Sci.* **15,** 1987–2001 (2006).
32. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157,** 105–132 (1982).
33. Wimley, W. C., Creamer, T. P. & White, S. H. Solvation energies of amino acid sidechains and backbone in a family of host–guest pentapeptides. *Biochemistry* **35,** 5109–5124 (1996).

ARTICLES

Chromatin remodelling at promoters suppresses antisense transcription

Iestyn Whitehouse[1], Oliver J. Rando[3], Jeff Delrow[2] & Toshio Tsukiyama[1]

Chromatin allows the eukaryotic cell to package its DNA efficiently. To understand how chromatin structure is controlled across the *Saccharomyces cerevisiae* genome, we have investigated the role of the ATP-dependent chromatin remodelling complex Isw2 in positioning nucleosomes. We find that Isw2 functions adjacent to promoter regions where it repositions nucleosomes at the interface between genic and intergenic sequences. Nucleosome repositioning by Isw2 is directional and results in increased nucleosome occupancy of the intergenic region. Loss of Isw2 activity leads to inappropriate transcription, resulting in the generation of both coding and noncoding transcripts. Here we show that Isw2 repositions nucleosomes to enforce directionality on transcription by preventing transcription initiation from cryptic sites. Our analyses reveal how chromatin is organized on a global scale and advance our understanding of how transcription is regulated.

Nucleosomes are the basic repeating units of chromatin. They are composed of an octamer of histone proteins around which DNA is tightly wrapped[1]. The DNA contained within nucleosomes is less accessible than linker DNA; as a result, processes that rely on access to the genome are influenced profoundly by the positions of nucleosomes along DNA[2]. However, little is known about the factors that govern nucleosome positioning *in vivo*.

One factor governing nucleosome positioning is the intrinsic DNA sequence preference for the histone octamer, and sequences that favour and disfavour nucleosome assembly have been described[3–5]. Recent genomic studies have used intrinsic sequence preferences to predict nucleosome positions across a genome on the basis of DNA sequence[6,7] with modest success—*in vivo* nucleosome positions are predicted well for only a subset of genomic loci. This probably reflects the fact that a variety of protein factors also contribute to nucleosome positioning *in vivo*. Principal among these are ATP-dependent chromatin remodelling enzymes that alter the positions of nucleosomes *in vivo* and *in vitro*[8]. Therefore, to understand how nucleosome positions are specified *in vivo*, it is important to understand not only the contribution made by the DNA sequence but also the role of factors such as chromatin remodelling enzymes[9].

Isw2 is one such ATP-dependent chromatin remodelling enzyme, and belongs to a family of proteins that are highly evolutionarily conserved[10,11]. In multicellular eukaryotes, Isw2 homologues have been implicated in the regulation of transcription[12,13], global chromosome structure[14], DNA replication[15,16], cell cycle progression[17], ribosomal DNA silencing[18,19] and cohesin loading[20]. In budding yeast, Isw2 acts as a gene repressor[21–23] by overriding the underlying nucleosome positioning signals of DNA[24], repositioning nucleosomes over unfavourable DNA sequences to establish a chromatin configuration that is repressive to transcription.

To understand global Isw2 function, we used high-resolution tiled microarrays to map the positions of nucleosomes across the yeast genome, and defined how their positioning is altered in a Δ*isw2* mutant strain. Nucleosome repositioning by Isw2 is directional and results in compact chromatin adjacent to sites of transcriptional activity. Loss of Isw2 leads to nucleosome positional changes and inappropriate transcription, resulting in the generation of noncoding transcripts. We find that chromatin remodelling is an important process that prevents aberrant transcription of the genome.

Genome-wide analysis of Isw2-dependent chromatin remodelling

We sought to discover Isw2 targets across the yeast genome by identifying nucleosomes whose positioning is altered in a Δ*isw2* mutant strain. We purified nucleosomal DNA from both wild-type and Δ*isw2* mutant yeast, and hybridized each to high-resolution tiled microarrays[25] that cover the entire yeast genome with ~5 base pair (bp) spacing (Supplementary Fig. 2).

Nucleosome-sized peaks were found across the genome; arrays of positioned nucleosomes generate a periodic signal whose maxima and minima are separated by the nucleosome repeat length of 165 bp. Comparison of wild-type nucleosome locations with those from a Δ*isw2* mutant revealed many sites of altered nucleosome positioning. To systematically identify these regions, we developed a comparative approach to detect differences in nucleosomal positioning and/or occupancy between wild-type and Δ*isw2* mutant strains (Supplementary Fig. 3). This analysis identified >1,000 distinct regions, typically of ~600 bp in length, where chromatin structure was disrupted in the Δ*isw2* mutant (Fig. 1 and Supplementary Table 1).

Having established the locations of chromatin changes, we used chromatin immunoprecipitation (ChIP) in conjunction with the

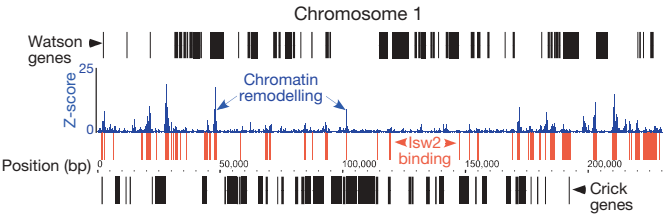


Figure 1 | Definition of Isw2 function using high-density microarrays. Shown is the global view of Isw2 binding and chromatin remodelling. Data for chromosome 1 are shown as an example. Black boxes represent ORFs, red boxes represent sites of Isw2(K215R) enrichment (Isw2 targets), and the blue line indicates sites of altered chromatin structure in an Δ*isw2* mutant.

[1]Division of Basic Sciences, [2]Genomics Resource, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA. [3]Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachusetts 01605, USA.

tiled microarrays to determine whether Isw2 was acting directly at these loci. Previously we found that the sites bound by a catalytically inactive mutant of the Isw2 protein, Isw2(K215R), are a reliable indicator of functional Isw2 targets, whereas the wild-type protein binds nonspecifically across the genome[26]. We therefore defined sites of Isw2 binding as sites at which Isw2(K215R) is enriched relative to the wild-type control. Figure 1 shows the sites of Isw2(K215R) binding and nucleosome positional changes across chromosome 1. Comparison of Isw2 localization with sites of Isw2-dependent nucleosome positional changes shows a tight association (Supplementary Fig. 4), suggesting that Isw2 is directly responsible for most of the chromatin changes we identify.

We then asked whether Isw2 is preferentially targeted to particular regions of the genome. Isw2 targets were divided into groups reflecting their proximity to genomic features (Supplementary Table 1). Isw2 function was detected at the 5′ ends of genes, as well as upstream of transfer RNA genes, consistent with previous reports[21,23,26,27]. In addition, we detected Isw2 activity at the 3′ end of genes and in intergenic regions distal to known genomic features. Examples of Isw2-dependent nucleosome repositioning at individual loci are shown in Supplementary Fig. 5.

Nucleosome repositioning at the 5′ end of genes

Isw2 is targeted to the 5′ end of ~20% of genes transcribed by RNA polymerase II, and Isw2-dependent chromatin remodelling was detected at ~35% of these targets (Supplementary Table 1). Because loss of Isw2 function at the 5′ end of genes is generally correlated with transcriptional derepression[21,23,27], chromatin structure changes mediated by Isw2 probably act to repress transcription at these sites. To define common features of Isw2-mediated nucleosome repositioning at the 5′ ends of genes, we aligned 5,767 non-dubious genes according to the midpoint of the first defined nucleosome (+1) upstream of the coding region of the gene (Fig. 2a). We first analysed the distribution of wild-type nucleosome positions, because previous high-resolution studies have only covered a fraction of the yeast genome[25,28]. Nucleosomes are highly organized at the 5′ end of genes and this organization spreads into the adjacent coding sequence. A short nucleosome-free region (NFR, which varies in size from gene to gene) typically lies upstream of nucleosome +1 (the first nucleosome upstream of the translation start site), and is the predominant site for transcription-factor-binding at promoters[25] (Fig. 2 and Supplementary Fig. 16b). The transcription start site is commonly downstream of the transcription-factor-binding sites and is generally occluded by the first half of nucleosome +1, consistent with recent analysis[28].

We then turned to the chromatin structure of Δ*isw2* mutants, asking if any general rules could be formulated for the action of Isw2 at the 5′ ends of genes. We selected genes that are enriched for Isw2 and display Isw2-dependent chromatin remodelling (~400 genes), and then aligned these loci according to the position of the +1 nucleosome determined for wild-type yeast. Next we overlaid the nucleosome hybridization signal from wild-type nucleosomes to generate an intensity map, the darkness of which is proportional to the number of traces at that coordinate; the average of the signal is shown as a green line (Fig. 3a). We generated a similar map for nucleosomes harvested from Δ*isw2* mutant yeast (Fig. 3b). To determine nucleosome repositioning trends at Isw2 targets, we calculated the difference between the wild-type and Δ*isw2* mutant intensity maps (Fig. 3c). This demonstrated a clear bias in nucleosome repositioning in the Δ*isw2* mutant strain. Specifically, there is a directional shift in the population of nucleosomes away from the NFR/intergenic region towards the adjacent genic sequence in Δ*isw2* mutants (Fig. 3c and Supplementary Fig. 6a, b). The size of the shift ranges up to ~70 bp, with a typical shift of ~15 bp that decreases in size for each successive nucleosome (Supplementary Fig. 7). Because loss of Isw2 function leads to a shift in nucleosomes away from the intergenic region, this finding implies that Isw2 functions to increase nucleosome occupancy at intergenic regions at the 5′end of genes by repositioning nucleosomes.

Nucleosome repositioning at the 3′end of genes

Nucleosome positional changes are also evident at the 3′ends of ~250 genes that are bound by the Isw2(K215R) protein (Supplementary Table1). We studied these targets by aligning the 3′ends of genes with respect to the nucleosome closest to the end of the open reading frame (ORF). We then compared how nucleosomes are repositioned at Isw2 targets and non-targets as described above. Again, we observe a distinct bias in the directionality of shifting; like at the 5′ end of genes, loss of Isw2 results in a shift of nucleosomes away from the intergenic region (Fig. 3f, and Supplementary Figs 6c, d and 7).

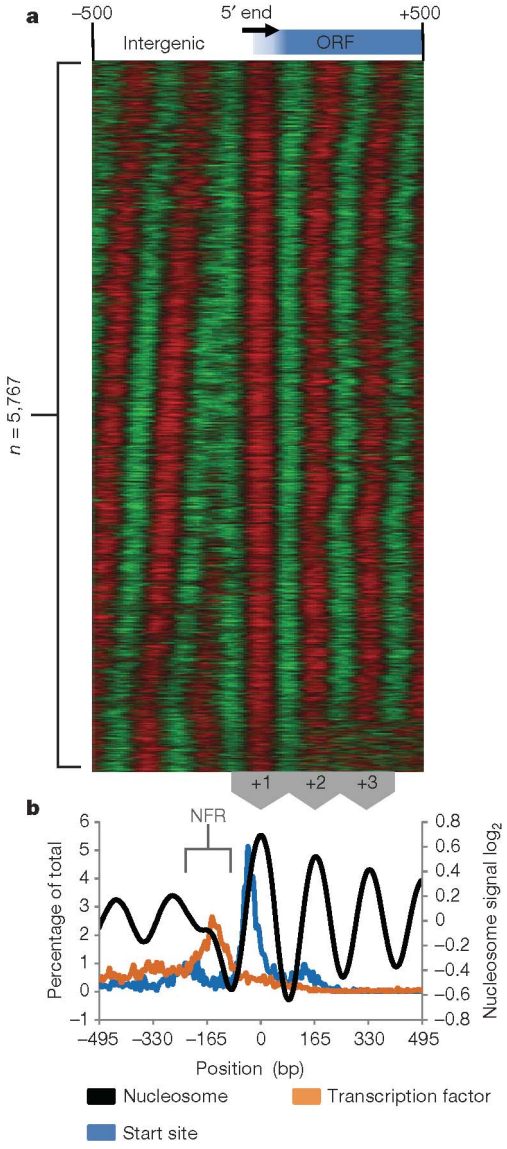


Figure 2 | Distinct nucleosome organization at the 5′ end of genes. **a**, A self-organizing map of nucleosome order at the 5′ end of genes. The 5′ ends of 5,767 non-dubious ORFs were aligned according to the +1 nucleosome adjacent to a short NFR. Loci with similar nucleosome arrangements are placed next to each other. Red represents a positive signal, whereas green is negative. **b**, The signal from **a** is averaged to illustrate typical nucleosome positions, shown as a black line. A frequency plot of the transcription start site from ref. 46 and the predicted transcription-factor-binding sites (with the most stringent cut-offs) from ref. 47 are displayed in blue and orange, respectively. Transcription factors typically bind within the NFR, whereas the transcription start site lies within the 5′end of the +1 nucleosome.

Thus, at both the 5′ and the 3′ ends of genes, Isw2 serves to shift nucleosomes onto adjacent intergenic regions.

Isw2 action at the 3′ ends of genes represents a major and previously unidentified class of targets. To characterize this group of targets more closely, we analysed their arrangement with respect to adjacent genomic features. In general, the intergenic region downstream of an ORF can be either convergent (containing the 3′ ends of two converging genes) or tandem (containing the 3′ end of the gene followed by the 5′ promoter of the adjacent gene). Our analysis revealed that more than 75% of the intergenic regions downstream of 3′ Isw2 targets are tandem, which is significantly higher than the genomic average of ~48%. Therefore, Isw2 activity at these targets seems to be correlated with the presence of an adjacent gene promoter. The high resolution of our analysis allows us to identify unambiguously the 3′ end of genes as a unique class of targets, rather than an artefact of Isw2 action at the 5′ end of the downstream gene (Fig. 4 and Supplementary Fig. 8).

Isw2 suppresses non-coding, antisense transcription

Because ~90% of all Isw2 targets are found adjacent to gene promoter regions, we considered a model in which Isw2 functions to repress transcription from promoters in general. According to this model, loss of Isw2 at the 5′ end of genes leads to increased transcription of the coding sequence. In contrast, loss of Isw2 action at the 3′ end of a gene would result in a shift in the nucleosome upstream of the NFR relative to the 5′ end of the adjacent gene. We reasoned that this might permit incorrectly oriented transcription to proceed from the adjacent promoter, resulting in the production of a noncoding, antisense transcript (Supplementary Fig. 1).

We sought direct evidence for the production of noncoding transcripts in yeast lacking Isw2. However, many noncoding transcripts are rapidly degraded by the exosome[29], complicating the search for these transcripts. A key component of this pathway is the poly(A) polymerase Trf4 (also known as Pap2); this polyadenylates transcripts, thereby targeting them for degradation[30–34]. We therefore generated a strain in which both *ISW2* and *TRF4* were deleted. The double-mutant strain displayed a synthetic slow growth phenotype that was more severe than either of the single mutants (Supplementary Fig. 9). This finding confirms previous results from high-throughput studies[35] and supports the idea that Isw2, like Trf4, has a critical role in the suppression of aberrant transcription.

To assay directly the presence of noncoding transcripts, we performed strand-specific northern analysis for a subset of genes in which Isw2 functions at the 3′ end. At each of the loci tested, we found that deletion of *ISW2* in a *trf4* background results in the generation of noncoding-antisense transcripts (Fig. 4a–c). At the gene *YGR166W* we found that production of a noncoding transcript is primarily dependent on the lack of Isw2 alone, whereas deletion of *TRF4* does influence transcript length (Fig. 4c). We used primer extension to determine the start site of the transcripts, and found that they are initiated at the 3′ end of the gene (Fig. 4 and Supplementary Fig. 10). We analysed the location of the antisense transcripts by use of high-resolution microarrays, and confirmed nucleosome repositioning by Southern blots (Supplementary Fig. 11). We also asked if noncoding transcripts are produced from other classes of Isw2 targets. At two loci—upstream of the tRNA gene *tT(UGU)G1*, and at an intergenic region, *iYDL025C*—we found evidence of noncoding transcription when both *ISW2* and *TRF4* are deleted (Supplementary Fig. 12). These results indicate a general role for Isw2 in the repression of noncoding transcription.

Discussion

Our studies provide a picture of how an ATP-dependent chromatin remodelling complex controls chromatin structure across a genome. We find that Isw2 functions at the interface between genic and intergenic regions, where it catalyses the directional shift of nucleosomes towards intergenic regions. Our data illustrate how nucleosomes are organized at regulatory sequences and how nucleosome repositioning is used to repress transcription from intergenic regions.

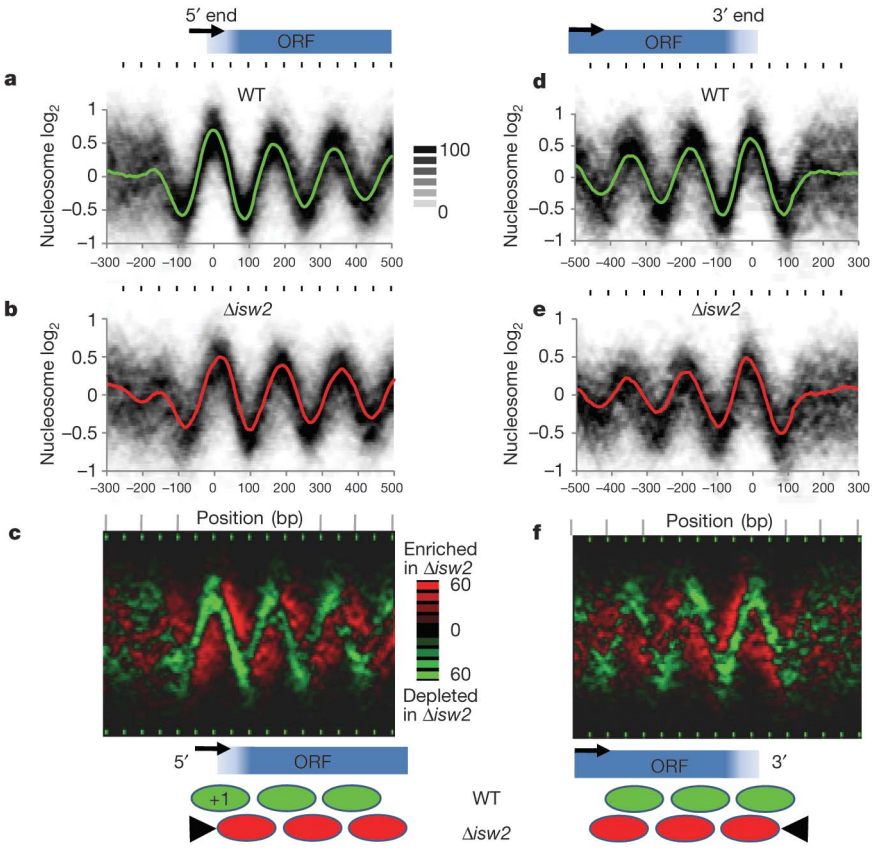


Figure 3 | Loss of Isw2 leads to directional nucleosome repositioning at the 5′ end and 3′ end of genes. **a**, **d**, Intensity maps of the nucleosomal signal at Isw2 targets at the 5′ end (**a**) and the 3′ end (**d**) of genes in wild-type (WT) cells. The average signal is shown as a green line. **b**, **e**, This is as described in **a** and **d** except using data from Δ*isw2* mutant yeast; the average signal is shown as a red line. **c**, **f**, Intensity maps showing nucleosome positional changes at Isw2 targets. Data from **a** were subtracted from that of **b** to generate **c**. Data from **d** were subtracted from that of **e** to generate **f**. Green represents regions that are depleted in an Δ*isw2* mutant strain, whereas red represents regions that are enriched. Nucleosome positional changes by Isw2 at 5′ and 3′ targets are illustrated at the bottom. The grey and colour scales represent the percentage of signal that is below that intensity. Black and green ticks are placed at 50-bp intervals.

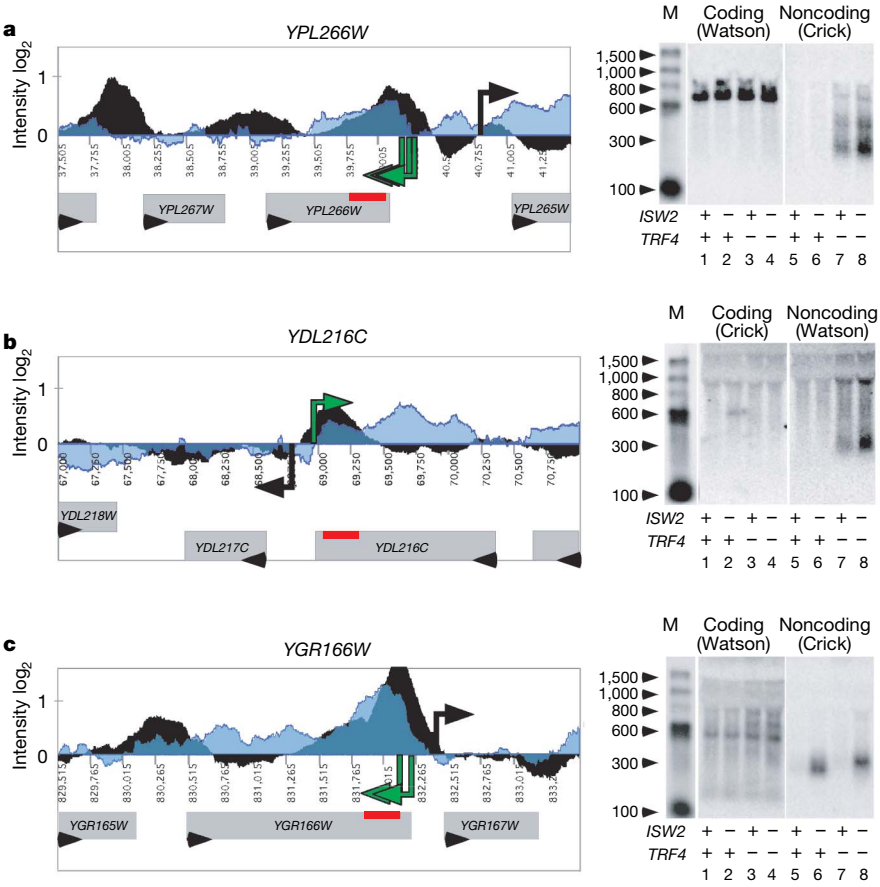


Figure 4 | Loss of Isw2 leads to noncoding transcription at the 3′ end of genes. Left, microarray data for Isw2 ChIP and RNA analyses. Log_2 ratio of the double mutant Δ*isw2* Δ*trf4* versus Δ*trf4* RNA signal is shown in transparent blue; Isw2(K215R) enrichment is shown in black. ORFs are shown as grey boxes, transcriptional direction is shown as a black arrow, and the location of northern blot probes used are shown as a red bar. Principal transcriptional start sites of the noncoding transcripts are shown as extended green arrows; start sites of coding transcripts mapped in ref. 48 are shown as extended black arrows. Right, strand-specific northern blots of Isw2 3′ targets. Loss of *ISW2* and *TRF4* results in the generation of noncoding-antisense transcripts at the genes *YPL266W* (**a**) and *YDR216C* (**b**); deletion of *ISW2* alone results in noncoding-antisense transcripts from the gene *YGR166W* (**c**). M, molecular weight marker.

We find that the positioning of thousands of nucleosomes adjacent to important regulatory sequences is controlled by Isw2. Yeast promoters frequently contain AT-rich DNA sequences[36] that have been found to inhibit nucleosome positioning[5,37,38]. Because Isw2 is able to use the energy from ATP hydrolysis to override the inherent nucleosome-positioning signal of the underlying DNA[24], Isw2 may function generally to reposition nucleosomes on unfavourable DNA sequences. Consistent with this, we find that poly dA/dT tracts, which are highly enriched at NFRs[25], are located within nucleosome +1 at many Isw2 targets (Supplementary Fig. 13). Loss of Isw2 would allow nucleosomes to adopt their inherent positioning preference, uncovering canonical or cryptic sites for transcriptional initiation. Because transcription is not necessary for the nucleosome positional changes caused by Isw2 deletion (Supplementary Fig. 14), transcription is likely to be a consequence rather than a cause of nucleosome repositioning at Isw2 targets. The broad scope of Isw2 action has implications for predictions of nucleosome positions on the basis of DNA sequence alone[6,7]. These studies have had some success, but at present they are unable to predict accurately many nucleosome positions within the cell[28] (Supplementary Fig. 15). The ability of proteins such as Isw2 to reposition nucleosomes provides a clear illustration that cellular factors actively operate to disrupt the intrinsic cues that would otherwise package the genome.

The primary site of action of Isw2 at the 5′ end of genes is the +1 nucleosome. This nucleosome is positioned such that the transcription start site is occluded by its 5′ edge[28] (Fig. 2). Upstream of +1 generally lies a short NFR, which typically contains transcription-factor-binding sites[25,28] (Fig. 2, Supplementary Fig. 16) and is probably the site for preinitiation-complex assembly. The nucleosome +1 generally contains the variant histone Htz1 (refs 28 and 39–41; Supplementary Fig. 17), which marks genes for rapid reactivation[42] and is subject to rapid replication-independent turnover[43]. This nucleosome is likely to act as a principal regulator of transcription, because RNA polymerase cannot reach the coding sequence without first transiting +1. In the context of this study, the specificity of Isw2 for this 'gatekeeper' nucleosome probably provides a regulatory mechanism to control gene expression by occluding the transcription start site or regulatory sequences through nucleosome repositioning (Supplementary Fig. 1).

A key finding of this study is that transcription is able to initiate from cryptic start sites when *ISW2* is deleted, which results in inappropriately oriented transcription from intergenic regions. This result is important because the mechanisms that ensure that transcription occurs in the correct orientation are largely unknown. Our findings suggest a model in which promoters are not intrinsically directional and can support inappropriately oriented transcription when chromatin structure is perturbed (Supplementary Fig. 1). Thus, transcription factors and DNA sequence alone are insufficient to prevent initiation from cryptic sites at these promoters. Because Isw2 remodels chromatin structure at the 3′ ends of many genes, the control of transcription by nucleosome positioning may be a general mechanism used by the cell.

METHODS SUMMARY

All yeast strains (S288C) were grown to mid-log phase ($D_{600} = 0.7$). Chromatin was crosslinked by the addition of formaldehyde, and was digested to mononucleosomes using micrococcal nuclease (MNase) and exonuclease III. Mononucleosomal DNA was purified by agarose gel electrophoresis. For microarrays, all samples were hybridized to *S. cerevisiae* tiling 1.0R arrays (Affymetrix). The signals from perfect match oligonucleotides were used to determine relative chromatin structure changes (Fig. 1, Supplementary Fig. 3); the difference between the hybridization signal from wild-type fragmented nucleosomal DNA and Δ*isw2* fragmented nucleosomal DNA was calculated. To determine nucleosome positions, the hybridization signal from fragmented nucleosomal DNA was normalized to the signal from full-length nucleosomal DNA. To align nucleosomes at the 5′ end of genes we first aligned all non-dubious ORFs according to the translation start site. These were then grouped using K-means clustering into ten nodes using Cluster[44]. Each node was manually aligned with respect to the nucleosome directly upstream of the NFR; if an

NFR was not apparent, we chose the first nucleosome upstream of the translation start site. This nucleosome was designated as +1. The data were further aligned by iteratively fitting an idealized nucleosome signal to the data. Signals with the best fit to the idealized nucleosome were defined as nucleosomes. Nucleosomes were organized at the 3′ end of genes in a similar manner to that at the 5′ ends. In this case the first nucleosome 5′ of the stop codon was chosen. For ChIP, all proteins were 3× Flag-tagged. Chromatin from Isw2 and Isw2(K215R) was fragmented using MNase as described in ref. 45; the immunoprecipitation method is described in ref. 26. Each immunoprecipitation sample is normalized to an input sample to control for experimental variations. For RNA analysis, yeast were grown to mid-log phase and the RNA was extracted with the use of acid phenol. Processed and raw data are available at http://www.fhcrc.org/science/labs/tsukiyama/supplemental_data/Global_Nuc_mapping.html.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 6 July; accepted 18 October 2007.

1. Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389,** 251–260 (1997).
2. Ehrenhofer-Murray, A. E. Chromatin dynamics at DNA replication, transcription and repair. *Eur. J. Biochem.* **271,** 2335–2349 (2004).
3. Satchwell, S. C., Drew, H. R. & Travers, A. A. Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* **191,** 659–675 (1986).
4. Lowary, P. T. & Widom, J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J. Mol. Biol.* **276,** 19–42 (1998).
5. Anderson, J. D. & Widom, J. Poly(dA-dT) promoter elements increase the equilibrium accessibility of nucleosomal DNA target sites. *Mol. Cell. Biol.* **21,** 3830–3839 (2001).
6. Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442,** 772–778 (2006).
7. Ioshikhes, I. P., Albert, I., Zanton, S. J. & Pugh, B. F. Nucleosome positions predicted through comparative genomics. *Nature Genet.* **38,** 1210–1215 (2006).
8. Narlikar, G. J., Fan, H. Y. & Kingston, R. E. Cooperation between complexes that regulate chromatin structure and transcription. *Cell* **108,** 475–487 (2002).
9. Rando, O. J. & Ahmad, K. Rules and regulation in the primary structure of chromatin. *Curr. Opin. Cell Biol.* **19,** 250–256 (2007).
10. Eisen, J. A., Sweder, K. S. & Hanawalt, P. C. Evolution of the SNF2 family of proteins: subfamilies with distinct sequences and functions. *Nucleic Acids Res.* **23,** 2715–2723 (1995).
11. Flaus, A., Martin, D. M., Barton, G. J. & Owen-Hughes, T. Identification of multiple distinct Snf2 subfamilies with conserved structural motifs. *Nucleic Acids Res.* **34,** 2887–2905 (2006).
12. Badenhorst, P., Voas, M., Rebay, I. & Wu, C. Biological functions of the ISWI chromatin remodeling complex NURF. *Genes Dev.* **16,** 3186–3198 (2002).
13. Yasui, D., Miyano, M., Cai, S., Varga-Weisz, P. & Kohwi-Shigematsu, T. SATB1 targets chromatin remodelling to regulate genes over long distances. *Nature* **419,** 641–645 (2002).
14. Deuring, R. *et al.* The ISWI chromatin-remodeling protein is required for gene expression and the maintenance of higher order chromatin structure *in vivo*. *Mol. Cell* **5,** 355–365 (2000).
15. Collins, N. *et al.* An ACF1–ISWI chromatin-remodeling complex is required for DNA replication through heterochromatin. *Nature Genet.* **32,** 627–632 (2002).
16. Poot, R. A. *et al.* The Williams syndrome transcription factor interacts with PCNA to target chromatin remodelling by ISWI to replication foci. *Nature Cell Biol.* **6,** 1236–1244 (2004).
17. Fyodorov, D. V., Blower, M. D., Karpen, G. H. & Kadonaga, J. T. Acf1 confers unique activities to ACF/CHRAC and promotes the formation rather than disruption of chromatin *in vivo*. *Genes Dev.* **18,** 170–183 (2004).
18. Zhou, Y., Santoro, R. & Grummt, I. The chromatin remodeling complex NoRC targets HDAC1 to the ribosomal gene promoter and represses RNA polymerase I transcription. *EMBO J.* **21,** 4632–4640 (2002).
19. Li, J., Langst, G. & Grummt, I. NoRC-dependent nucleosome positioning silences rRNA genes. *EMBO J.* **25,** 5735–5741 (2006).
20. Hakimi, M. A. *et al.* A chromatin remodelling complex that loads cohesin onto human chromosomes. *Nature* **418,** 994–998 (2002).
21. Goldmark, J. P., Fazzio, T. G., Estep, P. W., Church, G. M. & Tsukiyama, T. The Isw2 chromatin remodeling complex represses early meiotic genes upon recruitment by Ume6p. *Cell* **103,** 423–433 (2000).
22. Fazzio, T. G., Gelbart, M. E. & Tsukiyama, T. Two distinct mechanisms of chromatin interaction by the isw2 chromatin remodeling complex *in vivo*. *Mol. Cell. Biol.* **25,** 9165–9174 (2005).
23. Kent, N. A., Karabetsou, N., Politis, P. K. & Mellor, J. *In vivo* chromatin remodeling by yeast ISWI homologs Isw1p and Isw2p. *Genes Dev.* **15,** 619–626 (2001).
24. Whitehouse, I. & Tsukiyama, T. Antagonistic forces that position nucleosomes *in vivo*. *Nature Struct. Mol. Biol.* **13,** 633–640 (2006).
25. Yuan, G. C. *et al.* Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* **309,** 626–630 (2005).
26. Gelbart, M. E., Bachman, N., Delrow, J., Boeke, J. D. & Tsukiyama, T. Genome-wide identification of Isw2 chromatin-remodeling targets by localization of a catalytically inactive mutant. *Genes Dev.* **19,** 942–954 (2005).
27. Fazzio, T. G. *et al.* Widespread collaboration of Isw2 and Sin3–Rpd3 chromatin remodeling complexes in transcriptional repression. *Mol. Cell. Biol.* **21,** 6450–6460 (2001).
28. Albert, I. *et al.* Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature* **446,** 572–576 (2007).
29. Houseley, J., LaCava, J. & Tollervey, D. RNA-quality control by the exosome. *Nature Rev. Mol. Cell Biol.* **7,** 529–539 (2006).
30. Wyers, F. *et al.* Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. *Cell* **121,** 725–737 (2005).
31. LaCava, J. *et al.* RNA degradation by the exosome is promoted by a nuclear polyadenylation complex. *Cell* **121,** 713–724 (2005).
32. Vanacova, S. *et al.* A new yeast poly(A) polymerase complex involved in RNA quality control. *PLoS Biol.* **3,** e189 (2005).
33. Davis, C. A. & Ares, M. Jr. Accumulation of unstable promoter-associated transcripts upon loss of the nuclear exosome subunit Rrp6p in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **103,** 3262–3267 (2006).
34. Egecioglu, D. E., Henras, A. K. & Chanfreau, G. F. Contributions of Trf4p- and Trf5p-dependent polyadenylation to the processing and degradative functions of the yeast nuclear exosome. *RNA* **12,** 26–32 (2006).
35. Pan, X. *et al.* A DNA integrity network in the yeast *Saccharomyces cerevisiae*. *Cell* **124,** 1069–1081 (2006).
36. Behe, M. J. An overabundance of long oligopurine tracts occurs in the genome of simple and complex eukaryotes. *Nucleic Acids Res.* **23,** 689–695 (1995).
37. Kunkel, G. R. & Martinson, H. G. Nucleosomes will not form on double-stranded RNa or over poly(dA).poly(dT) tracts in recombinant DNA. *Nucleic Acids Res.* **9,** 6869–6888 (1981).
38. Iyer, V. & Struhl, K. Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J.* **14,** 2570–2579 (1995).
39. Raisner, R. M. *et al.* Histone variant H2A.Z marks the 5′ ends of both active and inactive genes in euchromatin. *Cell* **123,** 233–248 (2005).
40. Guillemette, B. *et al.* Variant histone H2A.Z is globally localized to the promoters of inactive yeast genes and regulates nucleosome positioning. *PLoS Biol.* **3,** e384 (2005).
41. Zhang, H., Roberts, D. N. & Cairns, B. R. Genome-wide dynamics of Htz1, a histone H2A variant that poises repressed/basal promoters for activation through histone loss. *Cell* **123,** 219–231 (2005).
42. Brickner, D. G. *et al.* H2A.Z-mediated localization of genes at the nuclear periphery confers epigenetic memory of previous transcriptional state. *PLoS Biol.* **5,** e81 (2007).
43. Dion, M. F. *et al.* Dynamics of replication-independent histone turnover in budding yeast. *Science* **315,** 1405–1408 (2007).
44. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95,** 14863–14868 (1998).
45. Liu, C. L. *et al.* Single-nucleosome mapping of histone modifications in *S. cerevisiae*. *PLoS Biol.* **3,** e328 (2005).
46. Zhang, Z. & Dietrich, F. S. Mapping of transcription start sites in *Saccharomyces cerevisiae* using 5′ SAGE. *Nucleic Acids Res.* **33,** 2838–2851 (2005).
47. MacIsaac, K. D. *et al.* An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics* **7,** 113 (2006).
48. Miura, F. *et al.* A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. *Proc. Natl Acad. Sci. USA* **103,** 17846–17851 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank members of the Tsukiyama and Henikoff laboratories for discussions, R. Basom for help with data analysis, and S. Henikoff, S. Biggins, S. Hahn and T. Owen-Hughes for critical reading of the manuscript. This work was supported by funds from NIGMS and the Leukemia and Lymphoma Society to T.T, from the Burroughs Wellcome Fund and Human Frontier Science Program to O.J.R, and from the NCI to J.D.

Author Contributions Experimental strategy was designed by I.W. and T.T., and experiments were performed by I.W. Preliminary nucleosome mapping was performed in collaboration with O.J.R. Data were analysed by I.W., with technical assistance from J.D. The paper was written by I.W., with assistance from T.T. All authors discussed the results and experiments, and edited the manuscript.

Author Information Raw data are deposited at GEO with accession numbers GSE8813, GSE8814 and GSE8815. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to T.T. (ttsukiya@fhcrc.org).

METHODS

Nucleosome harvest. Mononucleosomal DNA was prepared in a manner similar to that described in ref. 25 except nuclei were digested with the addition of ~700 units of MNase (Worthington) and 100 units of exonuclease III (NEB) for 20 min at 37 °C with constant agitation. Reactions were stopped with the addition of solution S (5% SDS, 100 mM EDTA). RNA and proteins were removed with RNaseA and proteinase K. Crosslinks were reversed by incubation at 65 °C for 16 h. SDS was precipitated from the solution with the addition of 3 M potassium acetate, pH 5.5 (to a final concentration of 750 mM), followed by centrifugation at 4,000*g*. DNA fragments were harvested from the supernatant using a G500 genomic column (Qiagen) according to instructions. DNA was then resolved on a 1.7% agarose gel, and the band corresponding to mononucleosomal DNA excised. Before labelling, full-length nucleosomal DNA fragments were treated with calf intestinal phosphatase (NEB) to remove 3′ phosphates.

Nucleosome data analysis. Nucleosome mapping data was normalized using quantile normalization, and was analysed using TAS software (version 1.1) provided by Affymetrix. We performed two sample analyses, in which there are two data sets termed a treatment and a control group. Each group consists of the subset of data falling within a bandwidth of 33 bp, resulting in n_t treatment probe intensities and n_c control probe intensities. Analysis at a particular position is based on all data aligning within ±1 bandwidth of the position. An estimate of fold enrichment of probes within the bandwidth is calculated using the Hodges–Lehmann estimator associated with the Wilcoxon rank-sum test.

Detection of Isw2-dependent chromatin remodelling. For global chromatin remodelling analysis (Fig. 1 and Supplementary Fig. 3), two biological replicates of wild-type fragmented nucleosomes were defined as the treatment group and two biological replicates of Δ*isw2* fragmented nucleosomes were defined as the control group using Tiling Analysis Software (TAS). The resulting $\log_2$ transformed signal was then spaced evenly at 5-bp intervals across the genome. Nucleosome-sized signals, S^n, were calculated by applying the following function at 5-bp intervals to the data set: $S_i^n = S_i(S_{i+165} + S_{i-165})$, where S corresponds to the signal at the chromosomal coordinate i. We defined sites of chromatin remodelling as 'blocks' whose signal is above 93% of the overall signal; this value was chosen because it gave the best correlation with Isw2(K215R) binding. Blocks less than 100 bp apart were classified as the same block.

Nucleosome signal normalization. To generate nucleosome signals, we initially fragmented the nucleosomal DNA into ~50-bp pieces and normalized its signal to that of DNaseI-digested genomic DNA. Nucleosome positions generated by this approach were highly consistent with previously published data (Supplementary Fig. 18). During our analysis, we found that nucleosome signals are highly dependent on the length of the DNA fragment that is hybridized to the microarray. Ends of full-length nucleosomal DNA fragments (~150 bp) from well-positioned nucleosomes hybridize with a ~2-fold greater efficiency than the mid points of the fragments. This is probably a steric effect caused by the relatively short length of oligonuclotides on the microarray. We find that the signal from these full-length nucleosomal DNA fragments provides an accurate means of identifying the ends of the nucleosomal DNA fragments (Supplementary Fig. 18). We found that nucleosome mapping could be improved significantly by generating a composite signal by using the data from fragmented nucleosomes to identify nucleosome peaks and by using the data from full-length nucleosomes to identify nucleosome edges. We found that normalizing the hybridization signal for fragmented nucleosomal DNA to that of full-length nucleosomal DNA generated an accurate nucleosome map that had better definition of nucleosomes than the signal generated from fragmented nucleosomes alone. Data normalized in this way were used in this paper; however, essentially the same results were obtained by normalizing the signals from fragmented nucleosomes to that of DNaseI-digested genomic DNA. Detrending was performed on the nucleosome hybridization data. This was done in a sliding window by subtracting the mean (of the maximum and minimum values over a 40 probe, ~200 bp, window) from the normalized data.

Nucleosome positional analysis. For nucleosome positional analysis, two biological replicates of wild-type fragmented nucleosomes were defined as the treatment group and two different biological replicates of wild-type full-length nucleosomes were defined as the control group. For nucleosome positioning in Δ*isw2* mutant yeast, two biological replicates of Δ*isw2* fragmented nucleosomes were defined as the treatment group and two different biological replicates of Δ*isw2* full-length nucleosomes were defined as the control group. Probe positions were set to the 3′ end of the oligonucleotide because this gave best correlation with published data sets. Nucleosome positions were determined by iteratively fitting an idealized nucleosome signal to a data set. The data set was first evenly spaced at 5-bp intervals and the probe position with the best fit (calculated by the Pearson correlation coefficient) to the idealized nucleosome signal was defined as the dyad. Figure 3 was created by aligning signals to a common point (+1 at the 5′ end or the nucleosome closest to the 3′ end of the gene), and then the frequency of nucleosome signal intensities was calculated at 5-bp intervals.

Generation of fragmented nucleosome DNA. Mononucleosomal DNA (3 μg) was placed in a 50 μl reaction containing 50 mM Tris:Cl, pH 6.8, 5 mM $MgCl_2$ and 0.3 μg μl^{-1} random hexamers. The reaction was incubated at 95 °C for 5 min, and was then chilled rapidly on ice. Five microlitres of dNTP mix (1.2 mM dGTP, 1.2 mM dCTP, 1.2 mM dATP, 0.95 mM dTTP, 0.25 mM dUTP) along with 50 units of Klenow Exo– (NEB) was added, and the reaction was allowed to proceed at 22 °C for 10 min and then at 37 °C for 30 min. The reaction was then heated to 95 °C for 5 min and then chilled rapidly on ice. A further 50 units of Klenow Exo– was added and the reaction was incubated at 22 °C for 10 min and then at 37 °C for 30 min. The reaction was stopped by phenol:chloroform extraction, and unincorporated random hexamers were removed with a gel filtration spin column. The DNA was then fragmented with the use of the wild-type Terminal Labelling Kit (Affymetrix).

Labelling of DNA. Samples to be hybridized on the microarrays (with the exception of full-length nucleosomal DNA fragments) were fragmented with the use of the wild-type Terminal Labelling Kit (Affymetrix) according to instructions. Complementary DNA was prepared and fragmented according to the wild-type double-stranded target assay manual for model organisms (Affymetrix). All DNA fragments were labelled using the wild-type Terminal Labelling Kit (Affymetrix) according to instructions.

ChIP analysis. Analysis of data from ChIP was performed using the Genomics Suite Software from Partek. Two biological replicates for wild-type Isw2 ChIP and Isw2(K215R) ChIP data sets were prepared. Each sample was normalized using quantile normalization. Initial signal intensity was calculated by normalizing the immunoprecipitation signal to that of the control input signal. The ratio of wild-type Isw2 input normalized signal to Isw2(K215R) input normalized signal was then calculated for each of the replicates. Regions of enrichment were identified using Partek software that identifies significant regions on the basis of a *t*-test on a sliding window (250 bp) centred around each probe; this tests the distribution of neighbouring values. Significant regions are defined as contiguous genomic regions that pass the *P*-value threshold ($P = 0.01$) in both samples. Only significant regions enriched in both replicates are used in this study.

RNA analysis. RNA samples hybridized to microarrays were normalized using quantile normalization and were analysed using TAS software (Affymetrix). Two-sample analysis was performed using a bandwidth of 66 bp. RNA harvested form Δ*trf4* yeast was set as the control group and RNA from Δ*trf4*Δ*isw2* was set as the treatment group. Primer extension reactions to map the transcription start sites were carried out using 30 μg of total RNA using the protocol described in ref. 26.

ARTICLES

The structural basis of calcium transport by the calcium pump

Claus Olesen[1,2], Martin Picard[3]†, Anne-Marie Lund Winther[1,3], Claus Gyrup[3], J. Preben Morth[1,3], Claus Oxvig[3], Jesper Vuust Møller[1,2] & Poul Nissen[1,3]

The sarcoplasmic reticulum Ca^{2+}-ATPase, a P-type ATPase, has a critical role in muscle function and metabolism. Here we present functional studies and three new crystal structures of the rabbit skeletal muscle Ca^{2+}-ATPase, representing the phosphoenzyme intermediates associated with Ca^{2+} binding, Ca^{2+} translocation and dephosphorylation, that are based on complexes with a functional ATP analogue, beryllium fluoride and aluminium fluoride, respectively. The structures complete the cycle of nucleotide binding and cation transport of Ca^{2+}-ATPase. Phosphorylation of the enzyme triggers the onset of a conformational change that leads to the opening of a luminal exit pathway defined by the transmembrane segments M1 through M6, which represent the canonical membrane domain of P-type pumps. Ca^{2+} release is promoted by translocation of the M4 helix, exposing Glu 309, Glu 771 and Asn 796 to the lumen. The mechanism explains how P-type ATPases are able to form the steep electrochemical gradients required for key functions in eukaryotic cells.

P-type cation pumps constitute an important family among actively transporting ATPases, with fundamental roles in cell function and in maintaining the cellular environment. Prominent examples include Ca^{2+}-ATPases and Na^+,K^+-ATPase, which alone consume approximately one-third of the energy used in humans[1]. Besides cation pumps, P-type ATPases are of critical importance in the homeostasis of heavy metals and the asymmetric distributions of lipids in membranes. The transport processes of the P-type ATPases are based on cyclical changes between two main conformational states, denoted E1 and E2, during which the ATPase is phosphorylated by ATP at a conserved aspartic acid side chain, and subsequently dephosphorylated. These processes are coupled to vectorial transport and counter transport by the controlled opening and closing of cytoplasmic and extracellular/luminal pathways.

Ca^{2+}-ATPases energize the Ca^{2+}-mediated signalling networks of the cell and maintain steep Ca^{2+} gradients across plasma membranes and inner membranes[2]. The sarco-endoplasmic reticulum Ca^{2+}-ATPase isoform 1a (SERCA1a) pumps Ca^{2+} from the cytosol of skeletal muscle cells into the sarcoplasmic reticulum store, thereby terminating a muscle contraction event[3]. SERCA1a activity results in transmembrane Ca^{2+} concentration ratios of three to four orders of magnitude (from a sub-micromolar to a millimolar level), in a high coupling ratio (approaching 2:1) of Ca^{2+} transport with ATP hydrolysis[4,5], and with counter transport of luminal protons[6–8].

Several aspects of the structure and overall transport mechanism of rabbit SERCA1a are well understood as a result of numerous investigations into the functional properties and owing to a range of crystal structures representing well-defined states of the functional cycle[9–15]. SERCA1a is composed of ten transmembrane helices (M1 through M10) and three cytosolic domains[9]: N (nucleotide binding), P (phosphorylation) and A (actuator). In its E1 form, SERCA1a binds two Ca^{2+} ions with high affinity in the middle of the ten transmembrane helices, whereas in the E2 form the cation-binding sites are associated with protons. Of particular importance, it has been shown that transition states of phosphorylation and dephosphorylation (as studied by aluminium fluoride complexes) are coupled to occlusion of bound calcium[11,12] and protons[14], respectively. In the forward direction of the functional cycle, the $Ca_2E1{\sim}P$ state, formed by reaction of Ca_2E1 with ATP, is converted to an E2P state, accompanied by translocation of Ca^{2+}. The E2P state is then dephosphorylated along with a counter transport of protons. All steps in the reaction cycle are reversible; thus ADP stimulates the dephosphorylation of the $Ca_2E1{\sim}P$ state through regeneration of ATP, and Ca^{2+}/Ca^{2+} exchange is observed at high vesicular levels of Ca^{2+} (refs 16–18) .

Despite detailed insight into the occluded transition states of phosphorylation and dephosphorylation, structural information on the essential step by which Ca^{2+} is transferred from the cytosol into the intraluminal space of sarcoplasmic reticulum vesicles is lacking. Does it occur through a gated channel (as, for example, suggested for Na^+,K^+-ATPase[19]) or is it the result of conformational changes that project the intramembrane binding sites towards the vesicle lumen[20]? In this article we address this question on the basis of new crystal structures of the Ca^{2+}-ATPase, all obtained in the absence of inhibitors bound in the membrane, namely: (1) the $Ca_2E1{\sim}P$ state formed by adenosine 5′-(β,γ-imido)-triphosphate (AMPPNP) and determined at 2.8 Å resolution, (2) an E2-BeF_3^- complex, representing a genuine E2P state, determined at 2.65 Å resolution, and (3) a structure determined at 3.0 Å resolution of the E2-AlF_4^- complex in the presence of the non-hydrolysable ATP analogue adenosine 5′-(β,γ-methylene)-triphosphate (AMPPCP), representing the ATP-modulated transition state of dephosphorylation, here denoted as E2-P*. We show how product separation in the high-energy $Ca_2E1{\sim}P$ state allows for the transition to the E2P state with exposure of a luminal pathway, which is resealed on the subsequent replacement of Ca^{2+} by H^+ and through formation of a dephosphorylation site in the E2-P* state. Together, these studies complete the overall view of SERCA1a structure and function and provide rationales that can be transferred to all members of the P-type ATPase family.

[1]Centre for Membrane Pumps in Cells and Disease—PUMPKIN, Danish National Research Foundation, and [2]Institute of Physiology and Biophysics, University of Aarhus, Ole Worms Alle, blg. 1185, DK - 8000 Aarhus C, Denmark. [3]Department of Molecular Biology, University of Aarhus, Gustav Wieds Vej 10C, DK - 8000 Aarhus C, Denmark. †Present address: Laboratoire de Cristallographie et RMN biologiques, UMR 8015 CNRS, Faculté de Pharmacie. Université Paris Descartes, 4 avenue de l'Observatoire, 75270 Paris Cedex 06, France.

Overall comparison of new structures

Crystals were obtained in the presence of native lipids and crystal structures were determined as described in Methods.

The three structures are shown in Fig. 1, along with the previously published structure of the ATP-bound E2 state[15] (using AMPPCP). Together, these structures represent a minimal scheme comprising four cornerstones of the functional cycle. Conformational changes occur as a result of concerted movements of transmembrane helices connecting the reactions at the phosphorylation site with binding events at the cation-binding sites in the membrane. The coupling is transmitted through movements of the A domain. Worth noting, E2-BeF_3^-– and E2-AlF_4^-–AMPPCP complexes are the first structures obtained of calcium-free SERCA1a without structural bias from membrane domain inhibitors like thapsigargin[10,14,15,21,22].

The Ca_2E1~P state

The conformation of SERCA1a crystallized in the presence of Ca^{2+} and AMPPNP is similar to the Ca_2E1-AMPPCP and Ca_2E1-ADP-AlF_4^- forms described earlier as the pre-state and transition state of phosphorylation, respectively, associated with occlusion of two Ca^{2+} ions in the membrane[11,12]. However, unbiased difference Fourier maps show important changes at the phosphorylation site (Fig. 2a), indicating that a genuine aspartyl-phosphoanhydride has formed at Asp 351 along with AMPPN (like ADP) at the nucleotide-binding pocket. The changes result in a ~3.1 Å net separation of the (bridging) β,γ-imido nitrogen and the γ-phosphoryl group transferred to Asp 351. The authenticity of phosphoenzyme formation was confirmed by mass spectrometry of re-dissolved crystals (Supplementary Fig. 1). In experiments on sarcoplasmic reticulum vesicles we also confirmed a previous finding[23] that SERCA1a cleaves AMPPNP in a Ca^{2+}-dependent reaction. The hydrolysis rate of AMPPNP (with a V_{max} of around 1 nmol per mg of protein per min at pH 7.1 and 23 °C) is approximately 5,000 times slower than observed with saturating amounts of ATP, yet Ca^{2+} transport is sustained with a high coupling ratio of 1.5 (±0.2):1 (Fig. 2b, c, and Supplementary Fig. 2), corresponding to what is observed with ATP at similar conditions[17,24]. In other words, AMPPNP is a fully functional but slow substrate for SERCA. The AMPPN (ADP) remains enclosed at the phosphorylation site with the aspartyl-phosphoanhydride trapped in a high-energy product complex, in accordance with the Ca_2E1~P state being ADP sensitive.

An important consequence of phosphorylation is that the nucleotide no longer bridges across the interface between the N and P domains: the N domain with the ADP-leaving group can therefore separate from the P domain, whereas the aspartyl-phosphoanhydride interacts with residues and a divalent cation of the P domain (Fig. 2a). Unlocking the domain interface by phosphorylation allows the ATPase to explore other conformational states that, according to recent kinetic evidence, lead to the rapid and reversible formation of E2P[25], as established previously in connection with the Na^+-dependent phosphorylation of Na^+,K^+-ATPase[26,27].

The E2P state with an exit pathway exposed to the lumen

The E2-BeF_3^- structure reveals a new and unique conformation of SERCA1a, showing an open, luminal pathway (Fig. 1, and Supplementary Fig. 4). The Asp 351 side chain is covalently modified by BeF_3^- and associated with a Mg^{2+} ion (Fig. 3a), and the site is isosteric to that of the phosphorylated Ca_2E1~P state. We find that

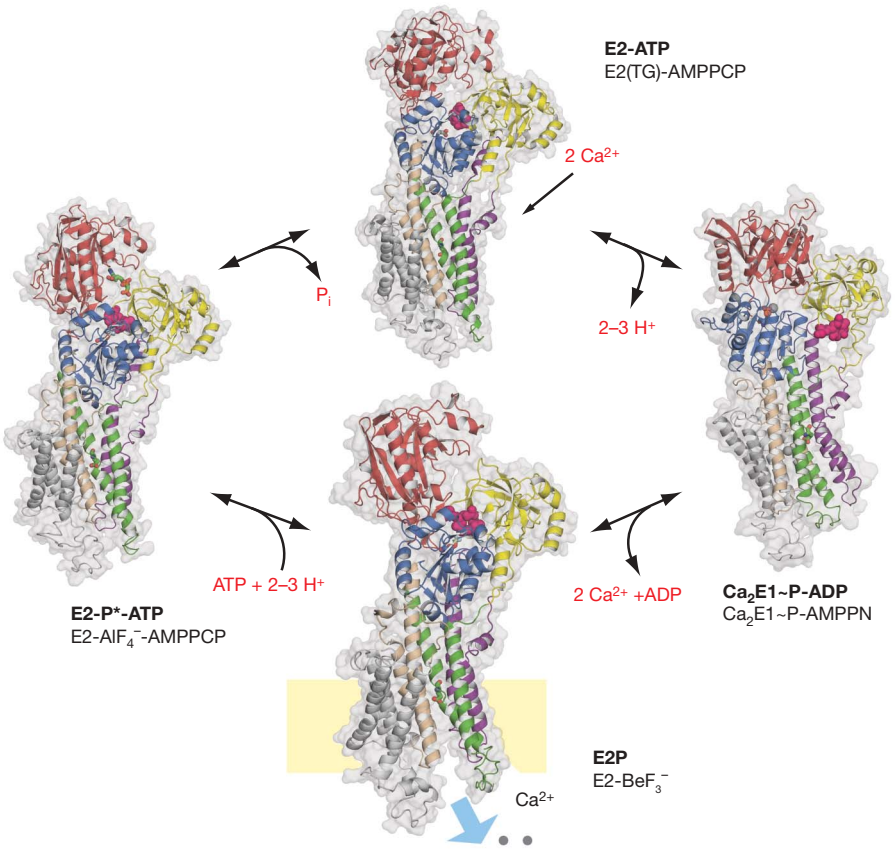


Figure 1 | Overall comparison of SERCA1a structures representing key states of the reaction cycle. The new structures of Ca_2E1~P-AMPPN, E2-BeF_3^- and E2-AlF_4^- complexes form the basis of this report and the E2-BeF_3^- complex is increased in size to emphasize its critical importance. Cation- and nucleotide-exchange reactions are indicated. The structures are depicted by grey, transparent surfaces and by cartoon representations, with the A domain in yellow, N domain in red, P domain in blue, transmembrane segment M1–2 in purple, M3–4 in green, M5–6 in wheat and M7–10 in grey. The TGES motif is shown by pink space-filling, residues 309, 771 and 796 (mentioned in the text) as sticks, and bound Ca^{2+} ions as grey spheres. Here, and in the following figures, structural representations were prepared with Pymol (http://pymol.sourceforge.net/).

BeF_3^- reacts with Ca^{2+}-ATPase with an affinity in the nanomolar range, yet our data also show that millimolar levels of Ca^{2+}, when granted access at the luminal side, lead to rapid and full re-activation of the BeF_3^--complexed Ca^{2+}-ATPase (Fig. 3b, c, and Supplementary Fig. 3), characteristic of an E2P state exposed to the lumen. Indeed, the signature by tryptophan fluorescence spectroscopy and a reduced Ca^{2+}-binding affinity have previously been used to argue that the BeF_3^--bound complex mimics the genuine E2P state[28,29].

A key question is then how phosphorylation has triggered the formation of an exit pathway? The answer is to be found in specific interactions of the A domain with the phosporylation site. The conserved TGES loop of the A domain replaces ADP and the N domain and forms a tight seal at the phosphorylation site by an extensive range of interactions with the Mg^{2+}-bound aspartyl phosphoanhydride (Fig. 3a, and see later). To yield this interaction of the TGES motif the A domain has rotated approximately 120° around the P domain and wedged itself into the N domain, which is now displaced from its interaction with the P domain. As a result of this movement the A domain exerts a powerful drag on transmembrane segments M1–M2 and M3–M4 and forces them to spread out and separate from M5–M6, which remain joined with the remaining transmembrane helices in an M5–M10 complex (Fig. 4). Indeed, intact linkers between the A domain and the membrane are of critical importance for Ca^{2+} transduction: thus proteolytic cleavage of the

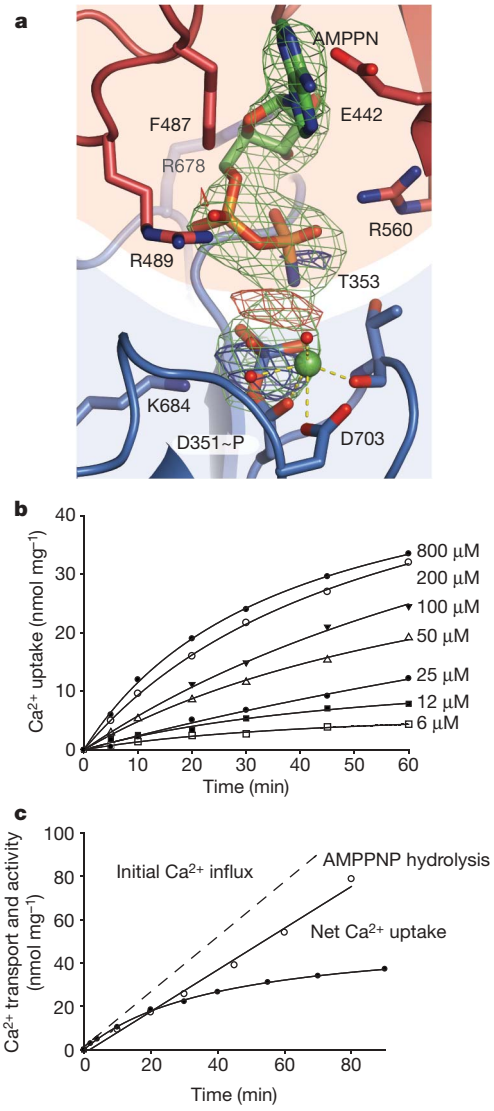


Figure 2 | The Ca2E1~P state obtained with AMPPNP. **a**, Refined structure (stick representation) of the Ca_2E1~P-AMPPN complex. The unbiased (F_O-F_C) difference map of the nucleotide complex (green mesh, 5.5σ), and the experimental ($F_O^{AMPPNP}-F_O^{ADP\text{-}AlF}$) difference map (−4σ and 4σ, red-brown and blue, respectively) show phosphoryl transfer to Asp 351. **b**, The upper panel shows Ca^{2+} accumulation by sarcoplasmic reticulum vesicles as a function of AMPPNP concentration. **c**, Ca^{2+} uptake, relative to AMPPNP hydrolysis at 200 μM of nucleotide. Dashed lines indicate initial rates of Ca^{2+} uptake, used for calculation of the coupling ratio (see text). The AMPPNP hydrolysis rate remains constant, whereas the Ca^{2+} transport rate declines owing to Ca^{2+}/Ca^{2+} exchange (J.V.M., unpublished observations).

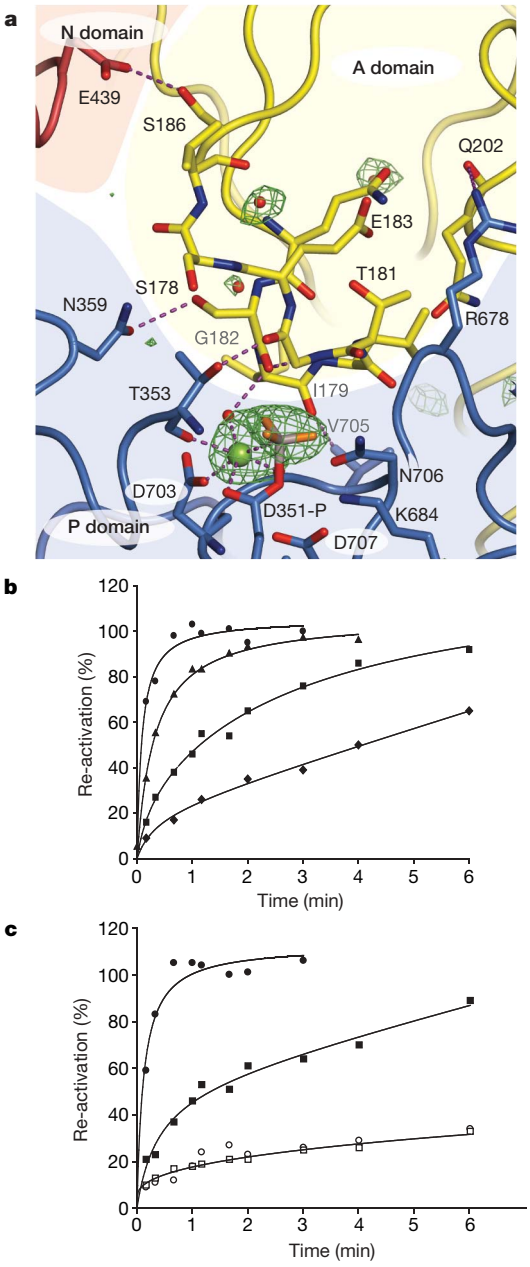


Figure 3 | The E2P state obtained with beryllium fluoride. **a**, Refined structure (stick representation) of the E2-BeF_3^- complex with unbiased (F_O-F_C) electron density shown at the phosphorylation site (5σ), showing BeF_3^- associated with Asp 351. Water molecules are shown as red spheres. **b**, **c**, Ca^{2+}-mediated displacement of BeF_3^- as a function Ca^{2+} concentration (circles, 10 mM; triangles, 5 mM; squares, 2 mM ; diamonds, 1 mM) using purified and leaky Ca^{2+}-ATPase preparations (**b**) or intact sarcoplasmic reticulum vesicles (**c**) in the presence (filled symbols) or absence (open symbols) of A23187 ionophore, thus distinguishing intravesicular and extravesicular effects of Ca^{2+}. *y* axis, recovery of activity relative to that of a BeF_3^- unmodified control; *x* axis, duration of Ca^{2+}-ATPase or vesicle exposure to 10 mM Ca^{2+}.

A-M3 linker[30] and shortening of the A-M1 linker[31] block formation of the E2P state, whereas proteolytic cleavage of the A-M2 linker[32] or extending the A-M1 linker[33] inhibit subsequent dephosphorylation of E2P, despite the conformational changes at the cytoplasmic domains. The transition of E1P—with a 10–15 Å hydrophobic barrier of the Ca^{2+}-binding sites to the lumen—to the E2P state leads to the formation of a three-lobed structure with a luminal pathway opening from the cation-binding sites towards the lumen (Fig. 4). The new configuration also distorts the calcium-binding sites in the membrane, primarily because of: (1) rotational movements of M6 bringing Asp 800 and Asn 796 away from their binding positions between Ca^{2+} sites I and II, (2) a ~6 Å translational movement of M4 with Glu 309 in the luminal direction, and (3) smaller translational movements of Glu 771 and Glu 908 (Fig. 4a, b). As a result, the luminal pathway exposes three of the Ca^{2+}-liganding residues, Glu 309 (site II), Glu 771 (site I) and Asn 796 (site I and II), to the lumen through a funnel-shaped polar pathway characterized by a width spanning from around 4 Å at the location of the exposed Ca^{2+}-liganding residues to around 15 Å at the luminal mouth of the pathway (Fig. 4d). In previous studies, exactly the same residues were pinpointed by site-directed mutagenesis as candidates for the luminal proton exchange during the reaction cycle[34,35]. The homologous residues in the occluded $K_2E2{\cdot}P_i$ state of Na^+,K^+-ATPase[36] (E327, E779, D804) interact with K^+ and are also important residues in formation of the E2P-like conformation of the palytoxin open channel[37].

The occluded E2-P* transition state without thapsigargin

The continuation of the functional cycle leads to the dephosphorylation reaction, as described earlier for the thapsigargin-bound complexes of SERCA1a with AlF_4^- mimicking the E2-P* transition state[14], or with MgF_4^{2-} mimicking the liberated but still entrapped phosphate of the $E2{\cdot}P_i$ state[13]. The authenticity of earlier E2 structures has been questioned, owing to the presence of thapsigargin or other inhibitors. However, the crystal structure of the E2-AlF_4^--AMPPCP complex obtained here in the absence of thapsigargin shows that the effect of the inhibitor is minimal (Supplementary Fig. 5), even at the binding pocket for thapsigargin between M3, M5 and M7. The inhibitory effect of thapsigargin thus seems to be based on trapping SERCA1a in the occluded E2 conformation. We have previously shown that the E2-AlF_4^- form represents an occluded state associated with the binding of probably 2–3 protons for counter transport[14], more correctly denoted H_{2-3}E2-AlF_4^-. Thus, during dephosphorylation, the M1–M2 and M3–M4 segments close against the M5–M10 region, burying the cation-binding residues in the membrane (Figs 1 and 4c). Worth noting, these residues adopt nearly the same configuration in thapsigargin (TG)-bound forms[10,13], and also in the E2-BeF_3^- complex, though stabilized by a Mg^{2+} ion in the latter (Fig. 4b, e). The reclosure of the transmembrane domain transmits to a ~10° perpendicular rotation of the A domain around the phosphorylation site of the P domain (away from the N domain and towards the membrane), in relation to which the TGES loop is moved by formation of a phosphorolytic site centred on Glu 183 (Fig. 5). The AMPPCP nucleotide of the E2-AlF_4^- complex (representing modulatory ATP) is bound at an exposed site centred on the Phe 487 region of the N domain and interacting with Lys 205 of the A domain (Fig. 5c), in a similar way to the binding of ADP to the E2(TG)-MgF_4^{2-} complex[13].

A nucleotide exchange mechanism of Ca^{2+}-ATPase

In transition from the Ca_2E1~P-ADP state to the E2P state the nucleotide-binding pocket at the N domain has gained access to the cytoplasmic environment, lined by the Arg 174 residue of the A domain (Fig. 5b). The E2-BeF_3^- structure (representing E2P) was obtained in the absence of nucleotide. However, AMPPN (representing the ADP leaving group) of the Ca_2E1~P-AMPPN nucleotide complex provides a snug fit in the E2-BeF_3^- structure when docked on the basis of a structural alignment of the N domain (Fig. 5b), and even when modelled as ATP. We suggest that this site provides the basis for ADP release in exchange for ATP in the E2P state. Putting this model into context with the structures of E2-AlF_4^--AMPPCP (this study), E2(TG)-AMPPCP[15], and the Ca_2E1~P-AMPPN form (this study) a systematic progression is observed during which ATP enters from an exchange site, moving to modulatory sites and then to

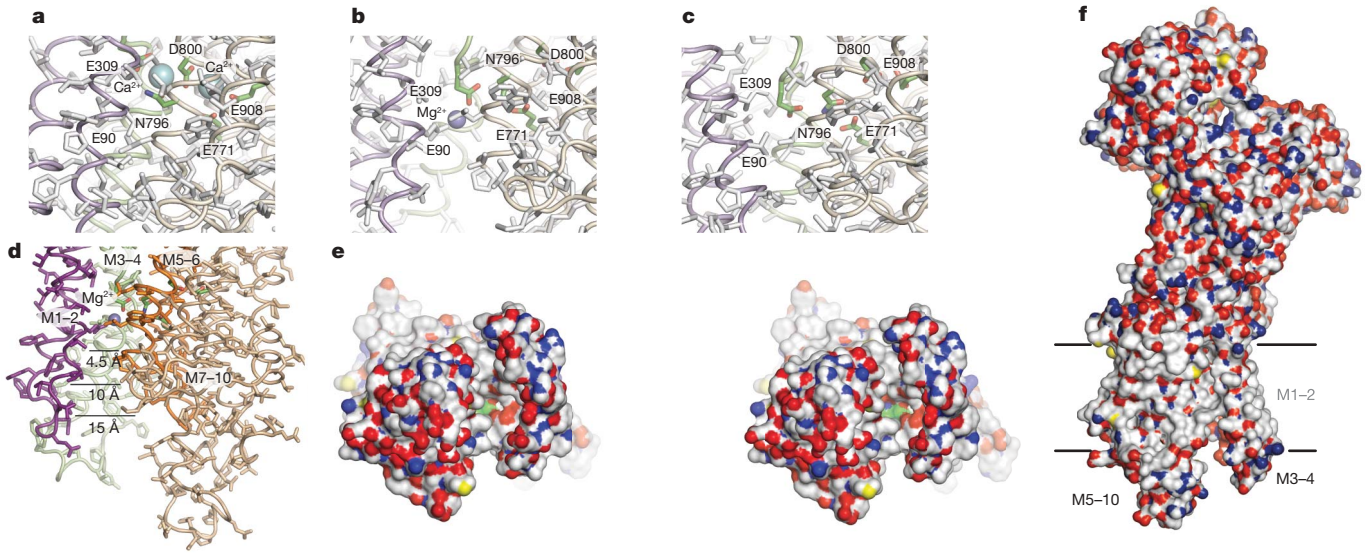


Figure 4 | The luminal exit pathway of sarcoplasmic reticulum Ca^{2+}-ATPase. **a**, The Ca_2E1~P form of SERCA1a (full stick representation and backbone trace) showing the Ca^{2+}-binding residues (carbon, green; oxygen, red; nitrogen, blue) and occluded Ca^{2+} ions (grey spheres). **b**, Same representation of the E2-BeF_3^- complex (grey-blue sphere for Mg^{2+}) with a similar orientation of the M7–10 region to **a**, now showing a luminal exit pathway and exposure of Ca^{2+} ligating residues. **c**, Similar representation of the E2-AlF_4^- structure with reclosure of the protonated state. **d**, Larger view of the E2-BeF_3^- form (as in **b**) with colours of transmembrane segments as in Fig. 1. Approximate dimensions of the funnel-shaped exit pathway are indicated. **e**, Stereoview of the E2-BeF_3^- complex (space-filling) as seen from the luminal side onto the membrane and showing a polar surface of the exit pathway. Carbon, white; sulphur, yellow; nitrogen, blue; oxygen, red, except for Ca^{2+}-binding residues, green (Glu 309, Glu 771 and Asn 796). **f**, Sideview of the E2-BeF_3^- complex (perpendicular to **e**), displaying well-defined boundaries of an apolar surface for membrane localization.

the final catalytic site as the enzyme proceeds from E2P (Fig. 5b) in the forward direction by Ca^{2+}-activated phosphorylation (Fig. 5). The Arg 174 residue of the A domain would seem to stabilize nucleotide binding in the E2P state. This residue is not conserved among P-type ATPases and may be the basis of the SERCA-specific ATP modulation of the $Ca_2E1{\sim}P$ to E2P transition[38].

Discussion

The present data indicate how two cytoplasmic Ca^{2+} ions, occluded in the membrane in the $Ca_2E1{\sim}P$ state, can be transferred to the lumen of the sarcoplasmic reticulum vesicles via an open luminal pathway formed by the separation of the M1–M2, M3–M4 and the M5–M6 segments in the E2P state. The transition also results in a decrease in Ca^{2+}-binding affinity owing to a change in the configuration of the liganding residues so that Ca^{2+} can be efficiently translocated from a low- to a high-concentration environment. The energy required is ultimately derived from ATP hydrolysis at the phosphorylation site—the motor of the pump—via a phosphoenzyme intermediate. All steps of the cycle are reversible, so how is efficient transport maintained as the electrochemical gradient builds up? The answer is to be found in the structural characteristics of the pump with the A domain linked to segments M1 through M3 to control the opening and closing of the luminal pathway, and the energetic input from the P domain connected to the membrane by the cytoplasmic extensions of M4 and M5. This provides the basis for coupling between the cation-binding sites in the membrane and the chemical processes occurring at the phosphorylation site. Furthermore, the physiological working conditions with a high ATP-to-ADP ratio in the cell have an important modulatory role, keeping the protein in a compact, nucleotide-bound state during the whole cycle (Figs 1 and 5).

To summarize, starting from the ATP-bound $H_{2-3}E2$ state (Figs 1 and 6), release of protons from the cation-liganding residue, in exchange for cytoplasmic Ca^{2+} ions, leads to the assembly of the phosphorylation site with the ATP-bound N domain[15], and with the A domain directing the occlusion of bound Ca^{2+} (refs 11 and 12). The $Ca_2E1{\sim}P$ state is formed through a favourable kinase activity producing ADP and an energy-rich aspartyl-phosphoanhydride bond, which unleashes the N domain from the P domain as the

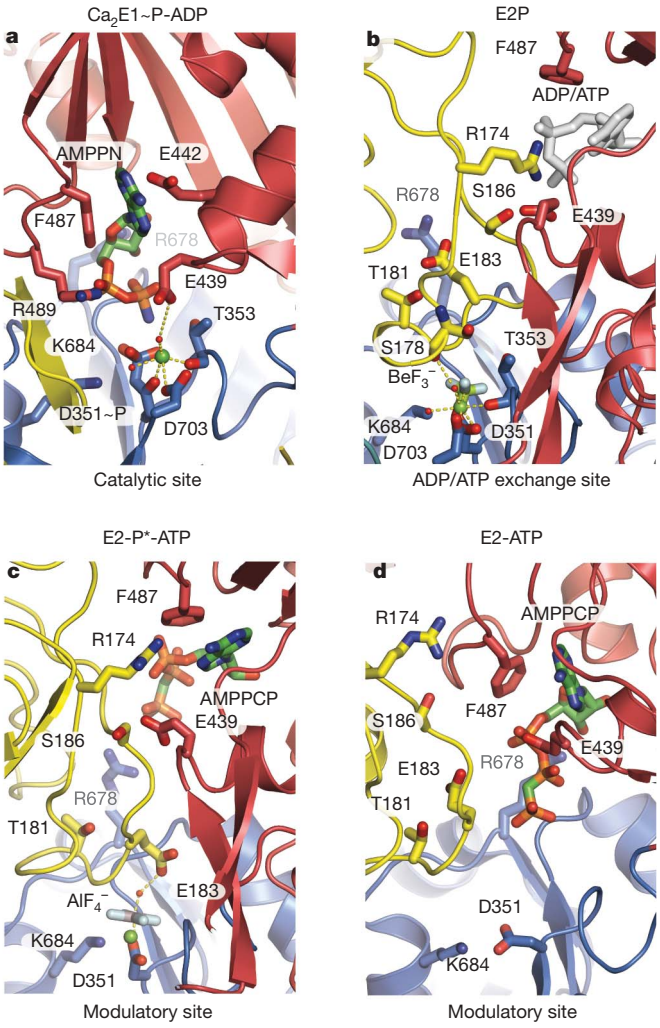


Figure 5 | Changes at the phosphorylation site of SERCA1a in the functional cycle. Structures are shown as a cartoon with selected residues in stick representation and with domain colours as in Fig. 1. The functional states and modes of nucleotide binding are indicated. **a**, $Ca_2E1{\sim}P$-AMPPN. **b**, E2-BeF_3^- structure with ADP (white stick) modelled at the N domain on the basis of the $Ca_2E1{\sim}P$-AMPPN structure. **c**, E2-$AlF4^-$-AMPPCP structure. **d**, E2-AMPPCP structure with thapsigargin (PDB 2C88). Functional state and mode of nucleotide binding are indicated on all panels.

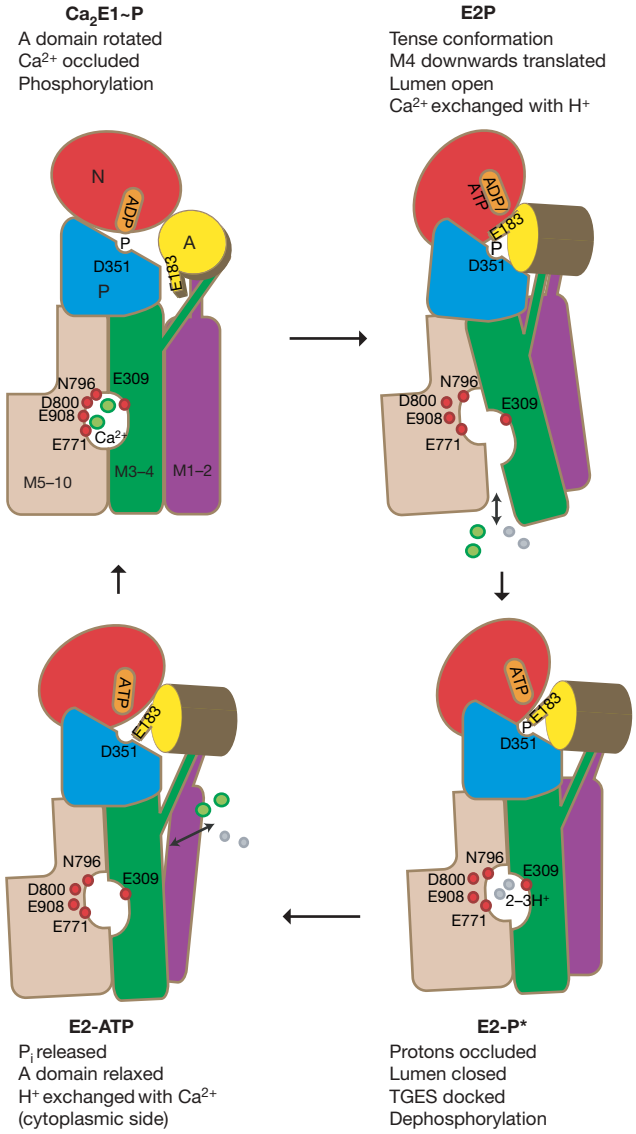


Figure 6 | Schematic representation of the reaction cycle. A schematic selection of key features of Ca^{2+}-ATPase function is indicated and reveals for the E2-P state the rotation of the A domain dragging M1–2, and the changes in the position of the P domain and N domain, pushing M3–4 in an outward and downward direction. See the Discussion for further detail. A domain, yellow; N domain, red; P domain, blue; helix M1–2, purple; M3–4, green; M5–10, wheat; Ca^{2+} ions and protons, green and grey spheres, respectively.

β,γ-phosphodiester bond breaks at the domain interface (Figs 2a and 5a). In the following step the A domain, with the conserved TGES loop, rotates towards the phosphorylation site, making firm associations with both the P domain and N domain. This movement exerts a downward push on the M3–M4 segment and a drag on the M1–M2 segment. In this transition, the configuration at the Ca^{2+}-binding residues in the membrane is forced apart, and the sites are favourably exposed to the polar environment of the luminal pathway to reduce the penalty of isolated charges in the membrane. During these movements, the cation-binding sites remain separated from the cytoplasm by a 15–20 Å thick barrier (Fig. 4b), whereas the N domain (left with ADP) is exposed to the cytosol and primed for ATP exchange (Fig. 5b), preventing a smooth reversal to the ADP-sensitive $Ca_2E1\sim P$ state.

Along with (partial) charge neutralization of the luminally exposed cation-binding sites by protons, a closure of the transmembrane segments becomes favourable, coupled to a downward rotation of the A domain and a movement of the P domain that leads the enzyme to the E2-P* transition state with occluded protons. In this state, the occluded protons are separated from the lumen by a 10–15 Å thick hydrophobic barrier (Fig. 4c) and a layer of positively charged residues at the luminal surface[14].

On dephosphorylation stimulated by the TGES motif, the cycle completes by release of the liberated phosphate stimulated by ATP binding[15] in the E2 state, with a cytoplasmic pathway opening for exchange of protons with calcium ions[10,15].

The pronounced conformational changes associated with Ca^{2+} translocation are in accordance with the role of Ca^{2+}-ATPase as a primary, active transporter. This is based on transitions between (at least) four different and well-defined conformational states, including the phosphoenzyme intermediates. A fundamental characteristic is a highly efficient separation between the cytoplasmic and extracellular/luminal environments at all steps in the cycle. The luminal pathway is broad and provides structural evidence for the long-sought E2P access channel on the basis of electrophysiological evidence[27,39] and from firmly anchored membrane transporters in general[40,41].

Glu 309 is a key player in Ca^{2+}-ATPase function, as previously pointed out[35,42] and can be described as a carrier of Ca^{2+} (and possibly protons) that moves up and down by about 6 Å with the M4 helix as the ATPase shuttles between E1 and E2 states (Fig. 6). Taken together, this pumping mechanism stands out as being powerful, and different to mechanisms based on narrow intramembranous pathways controlled by gating residues, such as those proposed for secondary transporters, like the lactose permease[43] and the sodium-coupled glutamate transporter[44]. At the same time, the coupled transport–counter-transport schemes are inherent to the system, compared to one-way transport of ABC transporters[45,46], and show a high resistance to backflow as compared to V-type ATPases; because of this, P-type ATPases perform well as electrogenic pumps.

The functional cycle of sarcoplasmic reticulum Ca^{2+}-ATPase rests on principles and structural elements that apply to all P-type ATPases. Indeed, P-type ATPases have been independently selected in fungi/plants (H^+-ATPase) and animals (Na^+,K^+-ATPase) for energization of strong plasma membrane potentials.

Finally, the open E2-BeF_3^- structure presented here provides a better template for modelling of the binding of clinically important inhibitors like omeprazole to H^+,K^+-ATPase (Supplementary Fig. 6), and ouabain to the Na^+,K^+-ATPase.

METHODS SUMMARY

Purification, crystallization and structure determination. Ca^{2+}-ATPase from rabbit fast-twitch skeletal muscle (SERCA1a) was prepared in sarcoplasmic reticulum membranes by differential centrifugations and solubilized by octaethyleneglycol dodecylether ($C_{12}E_8$) in buffers containing ligands as required for the formation of individual complexes. Supernatants from a final ultracentrifugation, with a protein concentration of approximately 12 mg ml^{-1}, were >90% pure and used directly for crystallization experiments. Crystals were obtained by the vapour diffusion method in hanging drops using PEG6000 or PEG2000 monomethyl-ether as the precipitant and *tert*-butanol, methylpentanediol (MPD), or dimethyl sulfoxide as additives.

Crystallographic data were collected at ESRF in Grenoble, France, and at EMBL/DESY in Hamburg (Supplementary Table 1). Structures were determined by molecular replacement (Supplementary Fig. 1) and the final model refinement employing the use of TLS parameterization was achieved with programs of the PHENIX package[47].

Detection of phosphorylated Ca^{2+}-ATPase by mass spectrometry. We developed a mass spectrometry method for detection of phosphorylation of Ca^{2+}-ATPase by AMPPNP, based on previous reports[48,49]. Fast hydrolysis of the aspartyl-phosphoanhydride residue prevents the direct detection by mass spectrometry. However, borohydrin reduction of the aspartyl-phosphoanhydride followed by CNBr treatment produces a Ca^{2+}-ATPase fragment (Met 327 to Met 361) with a stable homoserine occurring in place of phosphorylation at Asp 351. The peptide was identified and analysed by MALDI-TOF mass spectrometry (Supplementary Fig. 1).

Biochemical assays. The AMPPNP-supported Ca^{2+} transport by Ca^{2+}-ATPase-containing sarcoplasmic reticulum vesicles was assessed by current protocols[30,50] with particular care to avoid nucleotide contamination due to the low turn-over rate of AMPPNP-driven compared to ATP-driven transport (Supplementary Fig. 2). Ca^{2+} displacement of E2-BeF_3^- was followed by measuring the recovery of enzyme activity on Ca^{2+} addition; experiments using sarcoplasmic reticulum vesicles used the Ca^{2+} ionophore A23147 to differentiate between the effect of intravesicular and extravesicular Ca^{2+} (Supplementary Fig. 3).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 13 August; accepted 26 October 2007.

1. Rolfe, D. F. & Brown, G. C. Cellular energy utilization and molecular origin of standard metabolic rate in mammals. *Physiol. Rev.* **77,** 731–758 (1997).
2. Carafoli, E. Calcium signaling: a tale for all seasons. *Proc. Natl Acad. Sci. USA* **99,** 1115–1122 (2002).
3. Ebashi, S. & Lipmann, F. Adenosine triphosphate-linked concentration of calcium ions in a particulate fraction of rabbit muscle. *J. Cell Biol.* **14,** 389–400 (1962).
4. Hasselbach, W. Quantitative aspects of the calcium concept of excitation contraction coupling—a critical evaluation. *Basic Res. Cardiol.* **75,** 2–12 (1980).
5. De Meis, L. *The sarcoplasmic reticulum: Transport and Energy Transduction* (ed. Bittar, E. E.) (Wiley & Sons, New York, 1981).
6. Levy, D., Seigneuret, M., Bluzat, A. & Rigaud, J. L. Evidence for proton countertransport by the sarcoplasmic reticulum Ca^{2+}-ATPase during calcium transport in reconstituted proteoliposomes with low ionic permeability. *J. Biol. Chem.* **265,** 19524–19534 (1990).
7. Cornelius, F. & Moller, J. V. Electrogenic pump current of sarcoplasmic reticulum Ca^{2+}-ATPase reconstituted at high lipid/protein ratio. *FEBS Lett.* **284,** 46–50 (1991).
8. Yu, X., Carroll, S., Rigaud, J. L. & Inesi, G. H. + countertransport and electrogenicity of the sarcoplasmic reticulum Ca^{2+} pump in reconstituted proteoliposomes. *Biophys. J.* **64,** 1232–1242 (1993).
9. Toyoshima, C., Nakasako, M., Nomura, H. & Ogawa, H. Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 Å resolution. *Nature* **405,** 647–655 (2000).
10. Toyoshima, C. & Nomura, H. Structural changes in the calcium pump accompanying the dissociation of calcium. *Nature* **418,** 605–611 (2002).
11. Sorensen, T. L., Moller, J. V. & Nissen, P. Phosphoryl transfer and calcium ion occlusion in the calcium pump. *Science* **304,** 1672–1675 (2004).
12. Toyoshima, C. & Mizutani, T. Crystal structure of the calcium pump with a bound ATP analogue. *Nature* **430,** 529–535 (2004).
13. Toyoshima, C., Nomura, H. & Tsuda, T. Lumenal gating mechanism revealed in calcium pump crystal structures with phosphate analogues. *Nature* **432,** 361–368 (2004).
14. Olesen, C., Sorensen, T. L., Nielsen, R. C., Moller, J. V. & Nissen, P. Dephosphorylation of the calcium pump coupled to counterion occlusion. *Science* **306,** 2251–2255 (2004).
15. Jensen, A. M., Sorensen, T. L., Olesen, C., Moller, J. V. & Nissen, P. Modulatory and catalytic modes of ATP binding by the calcium pump. *EMBO J.* **25,** 2305–2314 (2006).
16. Feher, J. J. & Briggs, F. N. Determinants of calcium loading at steady state in sarcoplasmic reticulum. *Biochim. Biophys. Acta* **727,** 389–402 (1983).
17. Gerdes, U. & Moller, J. V. The Ca^{2+} permeability of sarcoplasmic reticulum vesicles. II. Ca^{2+} efflux in the energized state of the calcium pump. *Biochim. Biophys. Acta* **734,** 191–200 (1983).
18. Yu, X. & Inesi, G. Variable stoichiometric efficiency of Ca^{2+} and Sr^{2+} transport by the sarcoplasmic reticulum ATPase. *J. Biol. Chem.* **270,** 4361–4367 (1995).
19. Artigas, P. & Gadsby, D. C. Na+/K+-pump ligands modulate gating of palytoxin-induced ion channels. *Proc. Natl Acad. Sci. USA* **100,** 501–505 (2003).

20. Tanford, C. Translocation pathway in the catalysis of active transport. *Proc. Natl Acad. Sci. USA* **80,** 3701–3705 (1983).
21. Toyoshima, C. & Inesi, G. Structural basis of ion pumping by Ca^{2+}-ATPase of the sarcoplasmic reticulum. *Annu. Rev. Biochem.* **73,** 269–292 (2004).
22. Takahashi, M., Kondou, Y. & Toyoshima, C. Interdomain communication in calcium pump as revealed in the crystal structures with transmembrane inhibitors. *Proc. Natl Acad. Sci. USA* **104,** 5800–5805 (2007).
23. Taylor, J. S. Sarcoplasmic reticulum ATPase catalyzes hydrolysis of adenyl-5′-yl imidodiphosphate. *J. Biol. Chem.* **256,** 9793–9795 (1981).
24. Meltzer, S. & Berman, M. C. Effects of pH, temperature, and calcium concentration on the stoichiometry of the calcium pump of sarcoplasmic reticulum. *J. Biol. Chem.* **259,** 4244–4253 (1984).
25. Mahaney, J. E., Thomas, D. D. & Froehlich, J. P. The time-dependent distribution of phosphorylated intermediates in native sarcoplasmic reticulum Ca^{2+}-ATPase from skeletal muscle is not compatible with a linear kinetic model. *Biochemistry* **43,** 4400–4416 (2004).
26. Skou, J. C. The Na,K-pump. *Methods Enzymol.* **156,** 1–25 (1988).
27. Läuger, P. in *Electrogenic pumps* Ch. 8 (Sinauer Associates, Sunderland, Massachusetts, 1991).
28. Danko, S., Yamasaki, K., Daiho, T. & Suzuki, H. Distinct natures of beryllium fluoride-bound, aluminum fluoride-bound, and magnesium fluoride-bound stable analogues of an ADP-insensitive phosphoenzyme intermediate of sarcoplasmic reticulum Ca^{2+}-ATPase: changes in catalytic and transport sites during phosphoenzyme hydrolysis. *J. Biol. Chem.* **279,** 14991–14998 (2004).
29. Picard, M., Toyoshima, C. & Champeil, P. Effects of inhibitors on luminal opening of Ca^{2+} binding sites in an E2P-like complex of the sarcoplasmic reticulum Ca^{2+}-ATPase with Be^{2+}-fluoride. *J. Biol. Chem.* **281,** 3360–3369 (2006).
30. Moller, J. V. *et al.* Calcium transport by sarcoplasmic reticulum Ca^{2+}-ATPase. Role of the A domain and its C-terminal link with the transmembrane region. *J. Biol. Chem.* **277,** 38647–38659 (2002).
31. Daiho, T. *et al.* Deletions of any single residues in Glu^{40}–Ser^{48} loop connecting a domain and the first transmembrane helix of sarcoplasmic reticulum Ca^{2+}-ATPase result in almost complete inhibition of conformational transition and hydrolysis of phosphoenzyme intermediate. *J. Biol. Chem.* **278,** 39197–39204 (2003).
32. Lenoir, G. *et al.* Functional properties of sarcoplasmic reticulum Ca^{2+}-ATPase after proteolytic cleavage at Leu^{119}–Lys^{120}, close to the A-domain. *J. Biol. Chem.* **279,** 9156–9166 (2004).
33. Daiho, T., Yamasaki, K., Danko, S. & Suzuki, H. Critical role of Glu^{40}–Ser^{48} loop linking actuator domain and 1st transmembrane helix of Ca2+-ATPase in Ca2+ deocclusion and release from ADP-insensitive phosphoenzyme. *J. Biol. Chem.* **282,** 34429–34447 (2007).
34. Andersen, J. P. & Vilsen, B. Structure–function relationships of cation translocation by Ca^{2+}- and Na+, K+-ATPases studied by site-directed mutagenesis. *FEBS Lett.* **359,** 101–106 (1995).
35. Vilsen, B. & Andersen, J. P. Mutation to the glutamate in the fourth membrane segment of Na+,K+-ATPase and Ca^{2+}-ATPase affects cation binding from both sides of the membrane and destabilizes the occluded enzyme forms. *Biochemistry* **37,** 10961–10971 (1998).
36. Morth, J. P. *et al.* Crystal structure of the sodium–potassium pump. *Nature* doi:10.1038/nature06419 (this issue).
37. Reyes, N. & Gadsby, D. C. Ion permeation through the Na+,K+-ATPase. *Nature* **443,** 470–474 (2006).
38. Wakabayashi, S., Ogurusu, T. & Shigekawa, M. Factors influencing calcium release from the ADP-sensitive phosphoenzyme intermediate of the sarcoplasmic reticulum ATPase. *J. Biol. Chem.* **261,** 9762–9769 (1986).
39. Apell, H. J. Structure–function relationship in P-type ATPases–a biophysical approach. *Rev. Physiol. Biochem. Pharmacol.* **150,** 1–35 (2003).
40. Jardetzky, O. Simple allosteric model for membrane pumps. *Nature* **211,** 969–970 (1966).
41. Vidaver, G. A. Inhibition of parallel flux and augmentation of counter flux shown by transport models not involving a mobile carrier. *J. Theor. Biol.* **10,** 301–306 (1966).
42. Inesi, G., Ma, H., Lewis, D. & Xu, C. Ca^{2+} occlusion and gating function of Glu^{309} in the ADP-fluoroaluminate analog of the Ca^{2+}-ATPase phosphoenzyme intermediate. *J. Biol. Chem.* **279,** 31629–31637 (2004).
43. Abramson, J. *et al.* Structure and mechanism of the lactose permease of *Escherichia coli*. *Science* **301,** 610–615 (2003).
44. Boudker, O., Ryan, R. M., Yernool, D., Shimamoto, K. & Gouaux, E. Coupling substrate and ion binding to extracellular gate of a sodium-dependent aspartate transporter. *Nature* **445,** 387–393 (2007).
45. Dawson, R. J. & Locher, K. P. Structure of a bacterial multidrug ABC transporter. *Nature* **443,** 180–185 (2006).
46. Hvorup, R. N. *et al.* Asymmetry in the structure of the ABC transporter–binding protein complex BtuCD–BtuF. *Science* **317,** 1387–1390 (2006).
47. Adams, P. D. *et al.* PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D* **58,** 1948–1954 (2002).
48. Collet, J. F., Stroobant, V. & Van Schaftingen, E. Evidence for phosphotransferases phosphorylated on aspartate residue in N-terminal DXDX(T/V) motif. *Methods Enzymol.* **354,** 177–188 (2002).
49. Purich, D. L. Use of sodium borohydride to detect acyl-phosphate linkages in enzyme reactions. *Methods Enzymol.* **354,** 168–177 (2002).
50. Fiske, C. H. & Subbarow, Y. The colorimetric determination of phosphours. *J. Biol. Chem.* **26,** 375–400 (1925).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We dedicate this paper to the memory of B. Holm. We thank B. Nielsen, M.-B. Hemmingsen and A. M. Nielsen for technical assistance; J. L. Karlsen and F. Fredslund for technical discussions; and D. Flot and L. Gordon at beamlines ID 23-1 and -2 (operated jointly with EMBL-Grenoble) and ID 29 at the European Synchrotron Radiation Facility (ESRF) for help with data collection. Beamtime at the EMBL-DESY synchrotron Hamburg Germany is also acknowledged. This work was supported by the Danish Natural Science Research Council through the DANSYNC program, the Danish Medical Research Council, the Aarhus University Research Foundation, and the Novo Nordisk Foundation. C.Ol. is the recipient of a stipend from the PC Petersen Foundation and a PhD fellowship from the faculty of Health Sciences Aarhus University. A PhD fellowship (A.-M.L.W.) was financed by the Lundbeck Foundation. M.P. was supported by a post-doctoral fellowship from the Federation of European Biochemical Societies (FEBS) and P.N. is supported by a Hallas-Møller stipend of the Novo Nordisk Foundation.

Author Contributions C.Ol., M.P., A.-M.L.W., J.V.M. and P.N. contributed equally to this work. J.P.M. assisted with data collection and structure determination. C.Ox. and C.G. contributed with mass-spectrometry data and analysis.

Author Information The structural data have been deposited with the following codes in the Protein Data Bank: Ca2E1~P, 3BA6; E2-AlF4$^-$, 3B9R; and E2-BeF3$^-$, 3B9B. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to P.N. (pn@mb.au.dk) and J.V.M. (jvm@biophys.au.dk).

METHODS

Purification and solubilization. Ca^{2+}-ATPase prepared from rabbit fast twitch skeletal muscle (SERCA1a) was purified from sarcoplasmic reticulum vesicles according to established procedures[51].

$Ca_2E1\sim P$ crystals were obtained by solubilization with 30 mM octaethyleneglycol dodecylether ($C_{12}E_8$) in 100 mM MOPS-KOH, pH 6.8, 20% (w/v) glycerol, 80 mM KCl, 5 mM AMPPNP and 10 mM $CaCl_2$. The E2-BeF_3^- crystals were obtained by solubilization with 30 mM $C_{12}E_8$ in 100 mM MOPS-KOH, pH 7.0, 50 mM LiCl, 5 mM $MgCl_2$, 2 mM EGTA, 5 mM NaF and 0.66 mM $BeSO_4$. The E2-AlF_4^--AMPPCP complex was solubilized with 35 mM ($C_{12}E_8$) in 100 mM MOPS-KOH, pH 6.8, 20% glycerol, 80 mM KCl, 3 mM $MgCl_2$, 2 mM EGTA and 1 mM AMPPCP.

Crystallization. The solubilization was followed by ultracentrifugation at 4 °C for 35 min at 50,000 r.p.m. in a Beckman TLA-110 rotor. The protein was stored on ice overnight and then subjected again to ultracentrifugation for 15 min at 70,000 r.p.m. Supernatants, with a protein concentration of approximately 12 mg ml^{-1}, were > 95% pure and used directly for crystallization experiments. The hanging drop method was applied (at 19 °C) by mixing 2 µl of protein solution with 2 µl of crystallization buffer solution composed in the following way: $Ca_2E1\sim P$, 8% (w/v) PEG6000, 15% (w/v) glycerol, 200 mM sodium acetate, 4% (v/v) tert-butanol and 5 mM β-mercaptoethanol; E2-BeF_3^-, 18% (w/v) PEG 6000, 50 mM $MgSO_4$, 10% (v/v) glycerol and 4% DMSO; E2-AlF^--AMPPCP, 12% (w/v) PEG2000 mono-methylether, 200 mM $MgSO_4$ and 6% MPD.

Large, single crystals grew within a week for the $Ca_2E1\sim P$-AMPPN form and 2–3 weeks for the E2-BeF_3^- and E2-AlF_4^- forms. Crystals were taken directly from the mother liquor after limited dehydration ($Ca_2E1\sim P$-AMPPN) or on addition of glycerol to crystallization drops for cryo-protection (E2-BeF_3^- and E2-AlF_4^- crystals), mounted in nylon loops, and flash-cooled in liquid nitrogen, as described[52].

Data collection. Crystallographic data were collected at the beamlines ID23-2 and ID29 ESRF (European Synchrotron Radiation Facility) Grenoble, France, and EMBL/DESY (The EMBL X11 beamline at the DORIS storage ring, DESY) Hamburg, Germany. Data processing was performed with XDS[53], and phases obtained by molecular replacement using partial search models and the PHASER program[54]. The initial molecular replacement phases were refined by density modification through the RESOLVE prime-and-switch routine[55] (Supplementary Fig. 4) and later stages of model building were carried out on the basis of omit maps and $2F_O-F_C$ maps using O[56]. Refinement of structures was performed with the CNS program[57]. Final model refinement employing the use of TLS parameterization was achieved with programs of the PHENIX package[47]. The structures were validated using PROCHECK[58].

MALDI-TOF analysis of phosphoenzyme and crystal samples. Direct detection of the aspartyl-phosphoanhydride residue by mass spectrometry is not possible owing to nearly instant hydrolysis[59]. Instead we have used $NaBH_4$ to specifically reduce the aspartyl-phosphoanhydride into a stable homoserine residue[48,49,60–62]. Control samples of SERCA1a phosphoenzyme were made by incubating, on ice, sarcoplasmic reticulum vesicles at 0.2 mg of protein per ml in 50 mM MOPS (3-(*N*-morpholino)propanesulfonic acid) titrated with Tris (2-amino-2-hydroxymethyl-1,3-propanediol) at pH 7.8, 100 mM KCl, 5 mM $MgCl_2$, 50 µM $CaCl_2$ in the presence or in the absence of 5 µM ATP. The reactions were quenched by adding ice-cold 10% (w/v) trichloroacetic acid (TCA 10%) after 20 s of reaction. All samples were then kept on ice for 20 min and centrifuged for 20 min at 4,300 r.p.m. in a bench-top centrifuge (4 °C). Pellets were resuspended in 750 µl of ice-cold 5% TCA and centrifuged for 10 min at 72,000 r.p.m. in a Beckman TLA-110 rotor (at 4 °C). The pellets were finally rinsed in 10 mM ice-cold HCl, centrifuged 5 min at 72,000 r.p.m. and lyophilized.

Reduction of the aspartyl-phosphoanhydride to homoserine was performed by adding 100 µl of freshly prepared 10 µM $NaBH_4$ solution, suspended in 100 µl of DMSO, to the pellets. Incubation was performed for 20 min and quenched by addition of 200 µl of ice-cold 0.44 M $KClO_4$. After 30 min incubation on ice, the samples were centrifuged at 50,000 r.p.m. for 20 min in a Beckman TLA-110 rotor and the pellets lyophilized.

Crystals of the $Ca_2E1\sim P$-AMPPN form were harvested with crystal mounting loops and transferred to ice-cold TCA 10% with further manipulations, including $NaBH_4$ reduction, as described above.

Cyanogen Bromide (CNBr) fragmentation was performed by incubating the $NaBH_4$-reduced material with a 100-fold molar excess of CNBr over methionine. CNBr was dissolved in 70% trifluoroaceticacid (TFA) and the reactions were incubated overnight at room temperature. The patterns of fragmentation were compared for the different samples by matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometry (Supplementary Fig. 1). Mass spectra were acquired with a Voyager-DE PRO (MALDI-TOF) instrument (Applied Biosystems) operated in linear mode. Samples were prepared by crystallization of the peptide analytes with α-cyano-4-hydroxycinnamic acid. Final spectra were obtained by averaging 200 single shot spectra, and calibrated externally

AMPPNP-supported Ca^{2+} uptake by sarcoplasmic reticulum vesicles. Active Ca^{2+} transport by sarcoplasmic reticulum vesicles was measured by Millipore filtration with the aid of $^{45}Ca^{2+}$, according to current protocols but with the inclusion of some precautions arising from the slow rates of uptake sustained by the ATP analogue. The sarcoplasmic reticulum vesicles were kept suspended at a protein concentration of 0.2–1 mg ml^{-1} in a 30 mM imidazole buffer, pH 7.1, containing 150 mM NaCl, 5 mM Mg^{2+}, 0.10 mM $^{45}Ca^{2+}$ in the presence of various concentrations of AMPPNP (5–800 µM). Aliquots (comprising 200 µg of protein) were quenched at different time periods with 4 ml of unlabelled and ice-cold buffer, without added Ca^{2+}, and the vesicles were deposited on a 0.45 µm Millipore filter and further rinsed twice with 3 ml of ice-cold buffer. We found that when Ca^{2+} uptake measurements were performed in this way, there was a slow, time-dependent rise in Ca^{2+} retained by the filter, preceded by a jump in Ca^{2+} accumulation, occurring immediately after the addition of AMPPNP (Supplementary Fig. 2). This jump (the presence of which is also apparent from a previous report[23]) is not attributable to inefficient removal of Ca^{2+} bound on the outside of the vesicles, which shows that only small amounts of Ca^{2+} are retained on the filter in the absence of AMPPNP. Instead it is probably caused by rapid hydrolysis of ATP contaminating the AMPPNP reagent. In agreement with this view we found that the extent of the jump (which at a concentration of 0.8 mM nucleotide would correspond to a contamination of the order of 1.25–2.5%, assuming a coupling ratio between 1 and 2) was proportional to the concentration of AMPPNP. To remove contaminating ATP we routinely started the Ca^{2+} uptake experiments by pre-incubating the AMPPNP and $^{45}Ca^{2+}$ containing incubation medium with purified and leaky Ca^{2+}-ATPase membranes (at 0.050 mg of protein per ml) for 15 min, before starting the Ca^{2+} uptake experiments by the addition of sarcoplasmic reticulum vesicles. This protocol results in a regular uptake curve that eventually (after 90 min) leads to the same accumulation as observed without the pre-treatment with the leaky Ca^{2+}-ATPase membranes (Supplementary Fig. 2). From such curves, we could calculate the initial rates of Ca^{2+} transport after subtraction of a minor contribution (corresponding to 2–4 nmol of Ca^{2+} per mg of protein) that was measured immediately after the sarcoplasmic reticulum addition both in the presence and absence of AMPPNP and thus not attributable to Ca^{2+} transport.

Reaction of Ca^{2+}-ATPase with beryllium fluoride and Ca^{2+} reactivation. The use of beryllium fluoride (BeF_3^-) as a structural analogue of phosphorylation that forms an inhibitory complex with SERCA1a was pioneered by Murphy and Coll[63]. Intrinsic fluorescence and other evidence showing that complexation is compatible with the formation of an E2P state with luminally oriented Ca^{2+}-binding sites were later provided[28,29]. In our experiments, we find that in Ca^{2+}-depleted EGTA media Ca^{2+}-ATPase reacts with BeF_3^- by an irreversible reaction, characterized by apparent rate constants of 22 mM^{-1} min^{-1} at pH 7.2 and 13 mM^{-1} min^{-1} at pH 6.0, and eventually leading to complete inhibition of enzyme activity by low (micromolar) concentrations of BeF_3^-, provided that BeF_3^- is present in larger than stoichiometric amounts compared to SERCA1a (Supplementary Fig. 3). However, the stability of the complex is more sensitive to the presence of Ca^{2+} (Fig. 3) as compared to the E2-AlF_4^- complex[14,28]. This creates the problem when using enzymatic spectrophotometry[64] to follow BeF_3^- on- and off-reactions that unwanted re-activation readily occurs during the assay, resulting in downward deflecting curves (Supplementary Fig. 3). To minimize assay-induced re-activation we performed these measurements at low Ca^{2+} concentrations (at [Ca^{2+}]/[EGTA] concentration ratios of 0.9:1) and low ATP concentrations (0.1 mM instead of 5 mM MgATP), which resulted in nearly linear traces.

51. Andersen, J. P., Lassen, K. & Moller, J. V. Changes in Ca^{2+} affinity related to conformational transitions in the phosphorylated state of soluble monomeric Ca^{2+}-ATPase from sarcoplasmic reticulum. *J. Biol. Chem.* **260**, 371–380 (1985).
52. Sorensen, T. L., Olesen, C., Jensen, A. M., Moller, J. V. & Nissen, P. Crystals of sarcoplasmic reticulum Ca^{2+}-ATPase. *J. Biotechnol.* **124**, 704–716 (2006).
53. Kabsch, W. Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J. Appl. Cryst.* **26**, 795–800 (1993).
54. McCoy, A. J., Grosse-Kunstleve, R. W., Storoni, L. C. & Read, R. J. Likelihood-enhanced fast translation functions. *Acta Crystallogr. D* **61**, 458–464 (2005).
55. Terwilliger, T. C. Maximum-likelihood density modification. *Acta Crystallogr. D* **56**, 965–972 (2000).
56. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).

57. Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54,** 905–921 (1998).
58. Laskowski, R. A., Moss, D. S. & Thornton, J. M. Main-chain bond lengths and bond angles in protein structures. *J. Mol. Biol.* **231,** 1049–1067 (1993).
59. Sickmann, A. & Meyer, H. E. Phosphoamino acid analysis. *Proteomics* **1,** 200–206 (2001).
60. Allegrini, S. *et al.* Bovine cytosolic 5′-nucleotidase acts through the formation of an aspartate 52-phosphoenzyme intermediate. *J. Biol. Chem.* **276,** 33526–33532 (2001).
61. Collet, J. F., Stroobant, V. & Van Schaftingen, E. Mechanistic studies of phosphoserine phosphatase, an enzyme related to P-type ATPases. *J. Biol. Chem.* **274,** 33985–33990 (1999).
62. Sanders, D. A., Gillece-Castro, B. L., Stock, A. M., Burlingame, A. L. & Koshland, D. E. Jr. Identification of the site of phosphorylation of the chemotaxis response regulator protein, CheY. *J. Biol. Chem.* **264,** 21770–21778 (1989).
63. Murphy, A. J. & Coll, R. J. Fluoride is a slow, tight-binding inhibitor of the calcium ATPase of sarcoplasmic reticulum. *J. Biol. Chem.* **267,** 5229–5235 (1992).
64. Moller, J. V., Lind, K. E. & Andersen, J. P. Enzyme kinetics and substrate stabilization of detergent-solubilized and membraneous (Ca^{2+} + Mg^{2+})-activated ATPase from sarcoplasmic reticulum. Effect of protein–protein interactions. *J. Biol. Chem.* **255,** 1912–1920 (1980).

ARTICLES

Crystal structure of the sodium–potassium pump

J. Preben Morth[1,2], Bjørn P. Pedersen[1,2], Mads S. Toustrup-Jensen[1,3], Thomas L.-M. Sørensen[2]†, Janne Petersen[1,3], Jens Peter Andersen[1,3], Bente Vilsen[1,3]* & Poul Nissen[1,2]*

The Na^+,K^+-ATPase generates electrochemical gradients for sodium and potassium that are vital to animal cells, exchanging three sodium ions for two potassium ions across the plasma membrane during each cycle of ATP hydrolysis. Here we present the X-ray crystal structure at 3.5 Å resolution of the pig renal Na^+,K^+-ATPase with two rubidium ions bound (as potassium congeners) in an occluded state in the transmembrane part of the α-subunit. Several of the residues forming the cavity for rubidium/potassium occlusion in the Na^+,K^+-ATPase are homologous to those binding calcium in the Ca^{2+}-ATPase of sarco(endo)plasmic reticulum. The β- and γ-subunits specific to the Na^+,K^+-ATPase are associated with transmembrane helices αM7/αM10 and αM9, respectively. The γ-subunit corresponds to a fragment of the V-type ATPase c subunit. The carboxy terminus of the α-subunit is contained within a pocket between transmembrane helices and seems to be a novel regulatory element controlling sodium affinity, possibly influenced by the membrane potential.

The Na^+,K^+-ATPase, originally described in 1957 (ref. 1), is a membrane-bound ion pump belonging to the family of P-type ATPases. By using energy derived from ATP hydrolysis, the Na^+,K^+-ATPase generates electrochemical gradients for Na^+ and K^+ across the plasma membranes of animal cells, as required for electrical excitability, cellular uptake of ions, nutrients and neurotransmitters, and regulation of cell volume and intracellular pH. The transport is accomplished by enzyme conformational changes between two states, E1 and E2, that selectively bind three Na^+ and two K^+ ions, respectively (Fig. 1a); the ions become transiently 'occluded', that is, inaccessible to the medium on either side of the membrane[2,3]. The pump is sensitive to the membrane potential—the major voltage-dependent steps being associated with the binding and release of one of the three Na^+ ions[4,5].

The Na^+,K^+-ATPase consists of α- and β-subunits. The α-subunit contains the sites for binding of Na^+, K^+ and ATP and is homologous to single-subunit P-type ATPases like the Ca^{2+}-ATPase. The β-subunit is unique to the K^+-counter-transporting P-type ATPases, Na^+,K^+-ATPase and H^+,K^+-ATPase; it is required for routeing of the α-subunit to the plasma membrane and for occlusion of the K^+ ions[6,7]. A protein belonging to the FXYD family (γ-subunit in kidney outer medulla) is often associated with the αβ-complex as a third subunit and regulates the pumping activity in a tissue- and isoform-specific way[8,9].

Although the first crystal structure of a P-type ATPase, the sarco(endo)plasmic reticulum Ca^{2+}-ATPase (SERCA), appeared in 2000 (ref. 10) and was followed by structures of this enzyme in several conformations (for example, refs 11–13), the present work reveals the structure of a multi-subunit P-type ATPase, the pig renal Na^+,K^+-ATPase αβγ complex, with bound K^+/Rb^+ counterions.

Rb^+-occluded enzyme with bound magnesium fluoride

As a congener of K^+, Rb^+ is specifically recognized by the Na^+,K^+-ATPase and transported into the cell. In the E2 state Rb^+ is occluded[2,3,14], as indicated by a very slow dissociation of $^{86}Rb^+$ in the absence of ATP (bottom panel of Fig. 1b). A Rb^+-occluded enzyme ($[Rb_2]E2{\cdot}MgF_4^{2-}$) can also be formed in the presence of the phosphate analogue MgF_4^{2-} (Fig. 1b, middle panel). In the absence of MgF_4^{2-}, the de-occlusion of $^{86}Rb^+$ is markedly accelerated by ATP (ref. 2), whereas this effect is abolished by binding of the

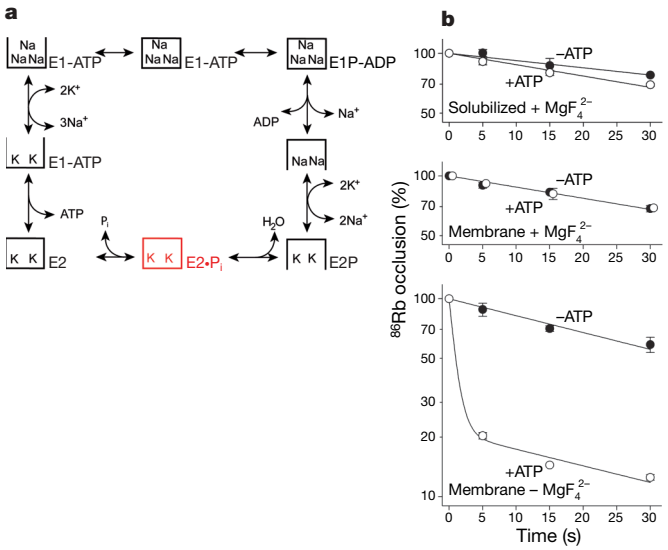


Figure 1 | Functional characterization of Na^+,K^+-ATPase used for crystallization. a, The reaction cycle of the Na^+,K^+-ATPase[2,3] showing in red the form crystallized. **b**, Demonstration of $^{86}Rb^+$ occlusion in the MgF_4^{2-}-bound form of the pig renal Na^+,K^+-ATPase from outer medulla. The dissociation of $^{86}Rb^+$ from membranous enzyme pre-incubated with $^{86}Rb^+$ in the absence (lowest panel) or presence (middle panel) of MgF_4^{2-}, and from $C_{12}E_8$-solubilized enzyme preincubated in the presence of MgF_4^{2-} (uppermost panel), was followed at 25 °C. Error bars, s.e.m.; $n = 3$.

[1]Centre for Membrane Pumps in Cells and Disease—PUMPKIN, Danish National Research Foundation, and [2]Department of Molecular Biology, University of Aarhus, Gustav Wieds Vej 10C, DK-8000 Aarhus C, Denmark. [3]Institute of Physiology and Biophysics, University of Aarhus, Ole Worms Allé, Blg. 1160, DK-8000 Aarhus C, Denmark. †Present address: Diamond Light Source Ltd, Diamond House, Chilton, Didcot, Oxfordshire, OX11 0DE, UK.
*These authors contributed equally to this work.

phosphate analogue. The MgF_4^{2-}-bound enzyme is expected to represent the $[K_2]E2{\cdot}P_i$ product state following dephosphorylation[15] (Fig. 1a). Hence, in the reaction sequence leading from the non-occluded K_2E2P state to the $[K_2]E2$ state the ions become occluded before the final P_i dissociation step, probably in relation to the hydrolysis of the covalent bond[13]. Following solubilization with the detergent $C_{12}E_8$, the occluded $[Rb_2]E2{\cdot}MgF_4^{2-}$ enzyme remained stable (Fig. 1b, top panel), and this form was used for crystallization.

Overall structure

Figure 2 and Fig. 3a present the crystal lattice and the overall architecture of the αβγ complex (see also Supplementary movie). The crystal lattice consists of layers of membrane-spanning regions stacked on each other (Fig. 2a). Between the membrane layers, the molecules are in contact through interactions between the cytoplasmic domains of molecules that are oriented head-to-head. The extracellular parts containing the glycosylated regions of the β-subunit do not contribute to the interlayer contact, but point into large solvent-filled channels of the crystal. The α-chain adopts a topology similar to that of the Ca^{2+}-ATPase with three characteristic cytoplasmic domains, the actuator (A), nucleotide-binding (N) and phosphorylation (P) domains, which together with all ten transmembrane segments, αM1–αM10, are well resolved in the electron-density maps (Fig. 2d). The model consists of α-subunit residues 19 to 1016 (complete C terminus, 998 residues), β-subunit residues 28 to 73 (46 residues, only the transmembrane segment), and a tentative assignment of γ-subunit residues 23 to 51 (29 residues, only the transmembrane segment). The asymmetric unit of the crystal lattice consists of two αβγ-units with limited contact between the A domains (Fig. 2b, c). There is no contact between the membrane parts of the α-subunits. The only membrane domain interaction occurs between β-subunits that are oppositely oriented relative to the membrane plane: an interaction that does not exist in the native membrane.

The occluded Rb^+/K^+-binding sites of the α-subunit

The anomalous scattering properties of rubidium allowed the accurate identification of two Rb^+ sites in the α-subunit by an anomalous difference Fourier map (Fig. 3b, and Table 1). With the enzyme having K^+ bound instead of Rb^+, it was possible under identical crystallization conditions to obtain crystals diffracting to 4 Å

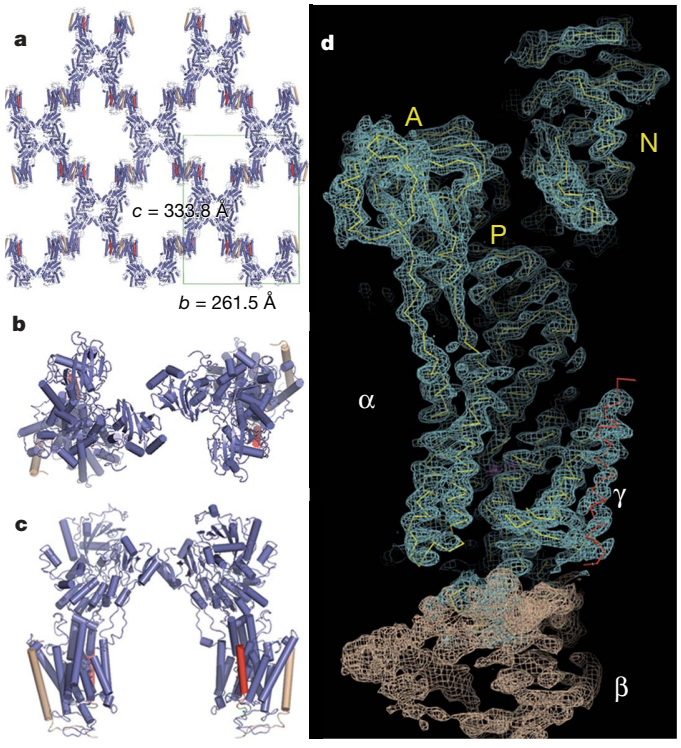


Figure 2 | Crystal packing and electron-density map. **a**, Crystal lattice viewed down the *a* axis. The unit cell is indicated. **b**, **c**, Asymmetric unit viewed down the *c* and *a* axis, respectively. The α-, β- and γ-subunits are coloured blue, wheat and red, respectively. **d**, Experimental electron-density map of the αβγ complex calculated at 3.5 Å and contoured at 1.0σ. The α- and γ-subunits are shown in cyan mesh (backbone indicated in yellow and red, respectively) and the extracellular part of the β-subunit as wheat mesh.

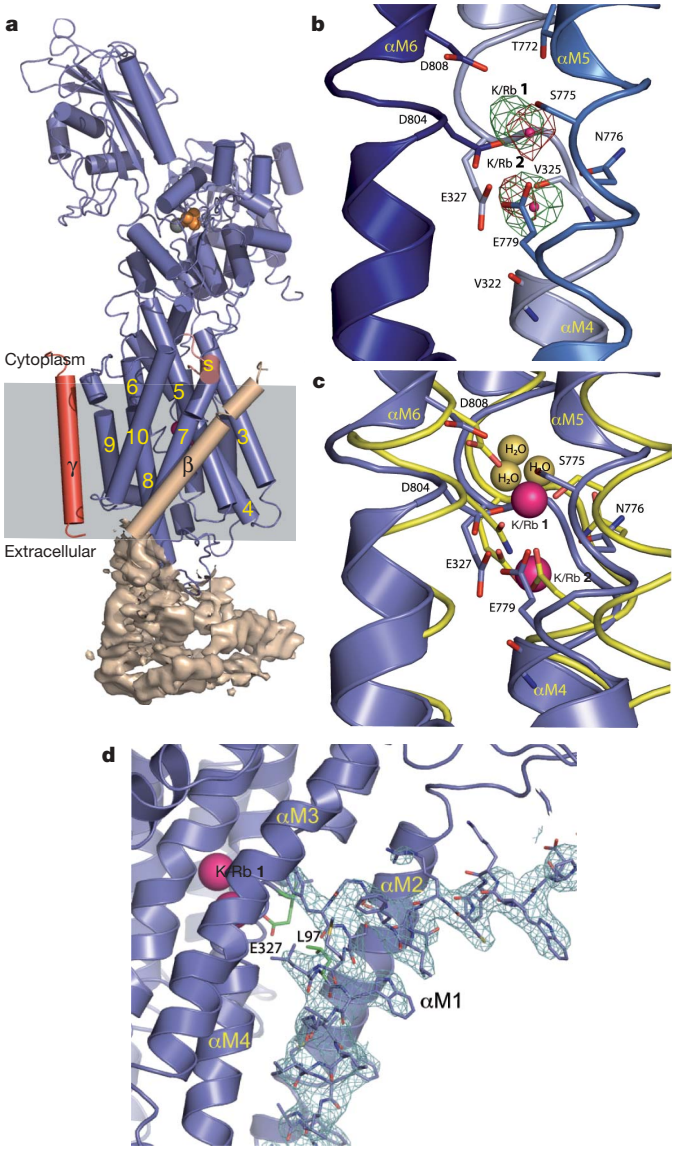


Figure 3 | Architecture of the Na^+,K^+-ATPase αβγ complex and the K^+/Rb^+ sites. The cytoplasmic side is up in all panels. **a**, The α-, β- and γ-subunits are coloured blue, wheat and red, respectively. Helices are represented by cylinders and β-strands by arrows. The β-ectodomain is shown by surface representation of the experimental electron density. The transmembrane segments of the α-subunit are numbered (yellow) starting with the most N-terminal. The small C-terminal helix (S, for switch) is light red. Mg^{2+}, MgF_4^{2-} and Rb^+ ions are grey, orange and pink, respectively. **b**, The red mesh (anomalous difference Fourier map) and the green mesh (omit F_O-F_C electron density map) show the positions of Rb^+ and K^+ ions, respectively. Oxygen-containing side chains within and close to the coordination sphere are shown. **c**, Structural alignment of the Na^+,K^+-ATPase (blue) with SERCA (yellow; PDB 1WPG) in the $E2{\cdot}MgF_4^{2-}$ forms. Yellow and magenta spheres represent water molecules in SERCA and K^+/Rb^+ ions in Na^+,K^+-ATPase, respectively. **d**, Interaction between Glu 327 (αM4) and Leu 97 (αM1)[17]. The cyan mesh indicates the electron-density map ($2F_O-F_C$) of αM1 contoured at 1.0σ.

Table 1 | Data collection, phasing and refinement statistics*

	Native $[Rb_2]E2{\cdot}MgF_4^{2-}$	Native $[K_2]E2{\cdot}MgF_4^{2-}$	Derivative $Ta_6Br_{12}^{2+}$ (ano.)	Derivative Orange-Pt (ano.)
Resolution (Å)	40–3.5 (3.6–3.5) †	40–4.0 (4.1–4.0)	48–6 (6.5–6.0)	48–7 (7.6–7.0)
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
Unit cell parameters *a*, *b* and *c* (Å)	68.9, 261.5, 333.8	69.3, 260.5, 334.6	69.2, 264.3, 337.4	70.0, 266.7, 338.0
R_{sym}‡ %	25.8 (>100)	20.9 (78.6)	5.6 (48.1)	9.9 (47.5)
Signal-to-noise ratio, $I/\sigma I$	10.8 (2.5)	7.5 (2.0)	10.3 (2.1)	8.5 (2.2)
Completeness (%)	99.8 (100)	95.9 (95.7)	96.3 (98.0)	98.4 (99.4)
Number of unique reflections	77,431	50,370	28,838	18,902
Redundancy	13.8	5.2	2.3	2.7
Phasing statistics				
Wave length (Å)	0.8146	0.979	1.078	1.073
Number of sites			6	8
Phasing power (iso./ano.)‡			1.30 / 1.01	0.85/0.41
Figure of merit (MIRAS; 40–60 Å)	0.40			
Figure of merit (DM; 40–3.5 Å)	0.78			

Refinement statistics		Number of atoms		Average B-factor (Å^2)		r.m.s. deviations	
Resolution (Å)	20–3.5	Subunits α/β/γ	15,480/740/456	Subunits α/β/γ	104/123/188	Bond lengths (Å)	0.008
R_{work}/R_{free}§ (%)	27.7/31.2	Ligands/ions	34	Ligands/ions	120	Bond angles (°)	1.51
						Ramachandran (%)‖	62.6/30.6/5.5/1.3

* For a full account of the data collection and structure determination see Methods.
† Values in parentheses here and below refer to the high-resolution shell as indicated.
‡ Phasing power is the root mean squared (r.m.s.) value of F_h divided by the r.m.s. lack-of-closure, as given by SHARP. Isomorphous and anomalous differences are given, respectively.
§ R_{free} is the *R*-factor calculated for a randomly picked subset with approximately 1,000 reflections excluded from the refinement throughout
‖Fractions of residues in 'most favourable', 'allowed', 'generously allowed' and 'disallowed' regions of the Ramachandran plot after refinement.

(Table 1). Two density peaks exceeding a 4σ level in the annealed omit map indicate the positions of the bound K^+ ions, which overlap with the Rb^+ sites (Fig. 3b). This confirms the expectation that K^+ and Rb^+ occupy similar sites. Notably, the Rb^+/K^+ ions are the first counter-ions directly visualized in a P-type ATPase structure. The two sites, here denoted 1 and 2, are found between the transmembrane helices αM4, αM5 and αM6. The Rb^+ ions are located in a common binding cavity, only ~4 Å apart with site 1 slightly closer to the cytoplasmic side of the membrane than site 2. No open pathways leading to the bound ions are apparent, in accordance with an occluded state. The side chains of residues Glu 327 (αM4), Ser 775, Asn 776, Glu 779 (αM5) and Asp 804 (αM6) are sufficiently close to the Rb^+ ions to donate ligands for binding (Fig. 3b), either directly or through an intervening water molecule. Asp 808 (αM6) is somewhat further away, but could be indirectly involved, and the same holds for Gln 923 (αM8) (see later and Supplementary Fig. 1). Asp 804 seems to donate a side-chain oxygen ligand to each Rb^+ ion (Fig. 3b). Glu 327 is associated exclusively with K^+/Rb^+ site 2 and may control the extracellular gate of the occlusion cavity[16,17], possibly guided by contact with Leu 97 of αM1 (ref. 17) (Fig. 3d). Most of these residues have been assigned a role in K^+ interaction by mutagenesis[16–21]. Identical residues are found at the corresponding positions in SERCA, except Asp 804 and Gln 923, which are replaced by asparagine and glutamate, respectively. In the E2P state of SERCA, Glu 309 (homologous to Glu 327) is exposed to the lumen for Ca^{2+} release[22]. The structure of the ion-binding cavity of SERCA in the E2 conformation that is supposed to accommodate the counter-transported protons seems rather similar to that of the K^+/Rb^+ cavity described here (Fig. 3c). A higher resolution would, however, be required to reveal subtle differences between the positions of the side chains in the Na^+,K^+-ATPase and the Ca^{2+}-ATPase. Notably, the water molecules identified in the higher-resolution $E2{\cdot}MgF_4^{2-}$ structure of SERCA[12] do not overlap directly with the Rb^+ ions observed in Na^+,K^+-ATPase (Fig. 3c).

The residues corresponding to Glu 327, Asn 776, Glu 779, Asp 804, Asp 808, and Gln 923 of the Na^+,K^+-ATPase all provide oxygen ligands for Ca^{2+} binding in the E1 form of SERCA[10,23,24], and they are therefore candidates for liganding residues in two of the three Na^+ sites in the E1 form of the Na^+,K^+-ATPase. If these residues do indeed bind the transported Na^+ ions, then a considerable overlap would seem to exist between the residues that coordinate K^+ in the E2 form and Na^+ in the E1 form, supporting the consecutive transport model[2]—in which Na^+ is released on the extracellular side in exchange for K^+ being transported to the cytoplasm through the same occlusion cavity—and, in more general terms, the alternating access model[25].

The β-subunit

The transmembrane helix of the β-subunit (βM) is clearly visible in the electron-density map (Fig. 4a). It traverses the membrane with a strong tilt of approximately 45° (Fig. 3a) and makes direct contact with αM7 and αM10 (Figs 3a and 4a). βM is closest to αM7, and approaches αM10 only near the extracellular end, in agreement with the finding that the β-subunit together with αM7 remains anchored in the membrane when αM8–M10 is released on heat denaturation[26]. Tyr 39, Phe 42 and Tyr 43 in βM are within interaction distance with αM7 residues around Gly 848, and the conserved glycines in the repeated GXXXG motif of βM (ref. 6) are exposed on the other side of βM (Fig. 4a). A dumb-bell-shaped density present between βM and αM7 may correspond to a phospholipid head group (Fig. 4a, 'PL').

The cytosolic amino-terminal part of the β-subunit cannot be modelled, but at low contour-level the density indicates that it continues around the α-subunit (Supplementary Fig. 2). The first 10–15 residues of the β-ectodomain have been tentatively traced and could come into contact with the αM7–αM8 loop around the SYGQ motif that is found to be crucial for αβ assembly[27]. Except for this part, it was not possible to build the β-ectodomain, although we find indications of an interleukin-receptor homology (Supplementary Fig. 3). However, in agreement with electron microscopy data[28,29], our density map provides a clear indication that the β-subunit completely covers the extracellular αM5–αM6 and αM7–αM8 loops as a lid (Figs 2d, 3a), which may relate to the essential role of the β-subunit in K^+ occlusion[7]. We interpret the disorder of the β-ectodomain as a direct consequence of its inherent flexibility in the absence of stabilizing crystal contacts.

The γ-subunit

The transmembrane segment of the γ-subunit (γM) is seen in the electron-density map as a stretch of approximately 30 amino acids with mostly α-helical structure (Fig. 4b). We noticed that the γ-subunit shares a sequence motif with the rotor ring c-subunit of

V-type ATPase from *Enterococcus hirae*[30] (Fig. 4c), possibly indicative of a common origin of these subunits. We used this analogy as a first resort in an assignment of γM, aided by the recent nuclear magnetic resonance model of the FXYD1 (ref. 31). The density maps further indicate that the extracellular part of the γ-subunit, containing the conserved FXYD motif, moves in between the α- and β-subunits where it may contact the β-subunit (Figs 2d and 4b). γM is clearly close to αM9 (Figs 3a and 4b), yet located on the outside of αM9 and not in the groove between αM9 and αM2, where it has been placed in modelling studies that are based on the Ca^{2+}-ATPase structure[8,32]. Several αM9 residues are within interaction distance of γM, including Phe 949, Glu 953, Leu 957 and Phe 960, in accordance with a mutagenesis study[32]. The part of the γ-chain showing the most intimate interaction with αM9 around Glu 953 contains Gly 41, which has been found mutated to arginine in familial dominant renal hypomagnesaemia[33].

Unique features of the α-subunit

As in the Ca^{2+}-ATPase, the αM4 and αM6 helices of the Na^+,K^+-ATPase are unwound in the middle, thereby making space for the ions (Fig. 3b), and αM1 shows a characteristic ~90° kink near the cytoplasmic surface of the membrane, where it comes into contact with αM3 (Fig. 3d). This contact point may function as a pivot for movement of αM1 in connection with ion binding[34]. The plant plasma membrane H^+-ATPase adopts a similar bent structure in M1 (ref. 35), suggesting that it constitutes a general structural motif of P-type ATPases.

Significant differences from SERCA are, on the other hand, seen in αM7, which is unwound at Gly 848, resulting in a kink, and at the cytoplasmic end of αM10 (Figs 4a and 5). These differences may be a consequence of the interaction with the β-subunit. Moreover, the C terminus may be influential (Fig. 5b, see further below).

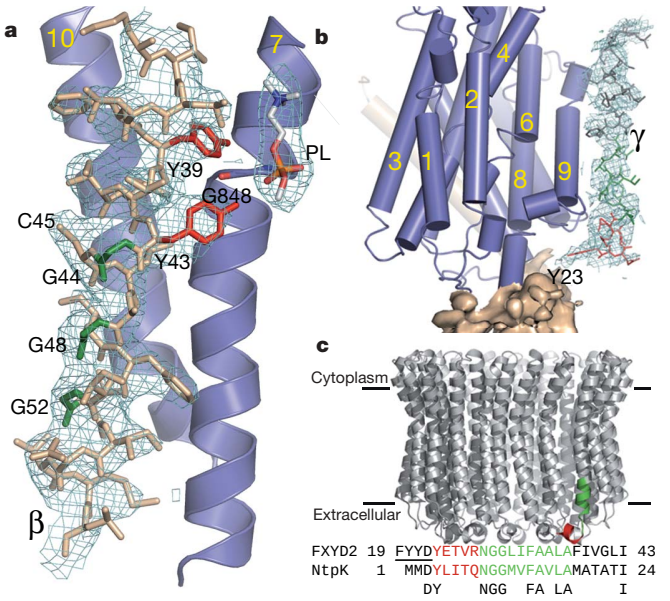


Figure 4 | Interactions between α and β and between α and γ transmembrane helices. **a**, βM (wheat colour) shown in stick representation. The electron-density map ($2F_O-F_C$) shown as cyan mesh was calculated at 3.5 Å and contoured at 1.0σ. PL, phospholipid head group modelled as phosphatidylcholine. **b**, γM represented by sticks and coloured according to the sequence alignment shown in **c**. The experimental electron density shown as cyan mesh is contoured at 1.0σ. Y23 indicates the start of the visible part of the γ-subunit. In **a** and **b** the transmembrane helices of the α-subunit are shown in blue with yellow numbering. **c**, The V-type ATPase rotor ring c subunit and sequence alignment with underlined signature sequence of the FXYD family. Structural elements showing sequence homology with the γ-subunit are coloured red and green.

The N domain is smaller than that of SERCA, which has insertions in surface loops, but is otherwise rather similar[28,36,37]. We find the N domain rather loosely associated with the rest of the molecule, rotated away from its interface with the A domain (Fig. 5a). The architectures of the A domain and P domain are also very similar in the two pumps, the MgF_4^{2-} in the catalytic site being coordinated by conserved residues from both of these domains. The C-terminal part of the P domain of the Na^+,K^+-ATPase contains a 20-residue-long outward-protruding insertion[28], which is seen to adopt the form of

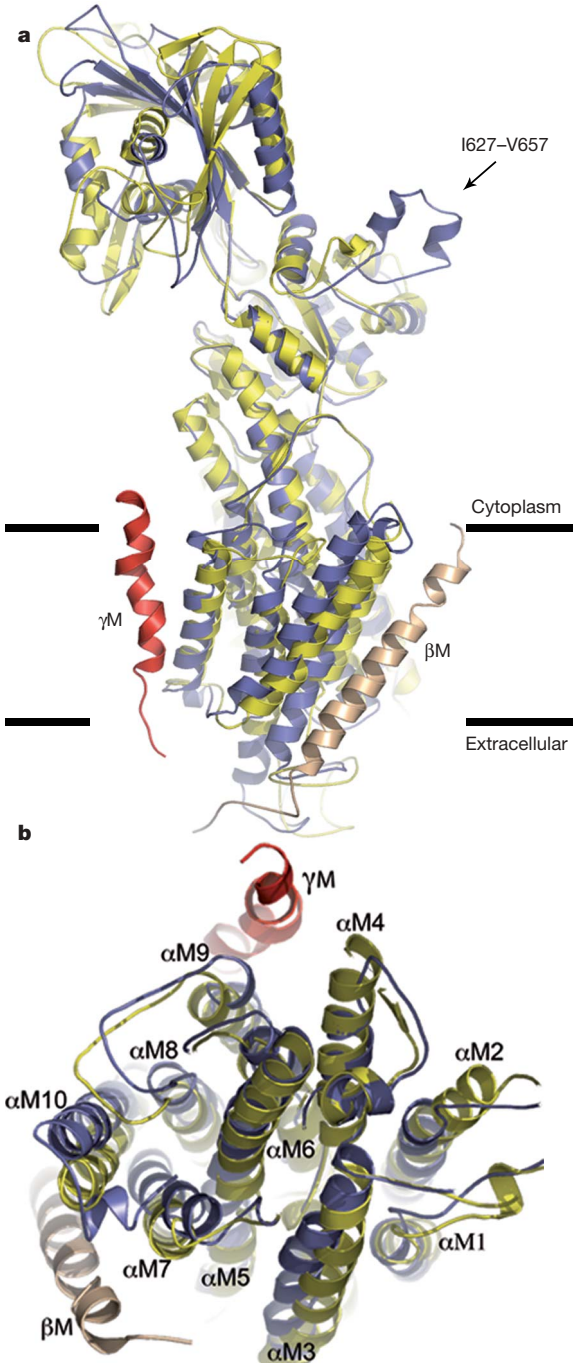


Figure 5 | Structural comparison of the Na^+,K^+ ATPase with the Ca^{2+}-ATPase. **a**, The Na^+,K^+ ATPase (blue) and the Ca^{2+}-ATPase (yellow) were structurally aligned using as fix points the highly similar A domain and P domain. In this view the A domain is not apparent. The arrow indicates a protrusion of the P domain unique to the Na^+,K^+ ATPase. **b**, Top view of the aligned transmembrane regions seen from the cytoplasmic surface. Note the triangular pocket between βM, αM7 and αM10 that accommodates the small helix of the C terminus of the α-subunit.

two small helices connected by a loop, as a possible target for interaction with regulatory proteins (Fig. 5a).

The C-terminal extension is crucial for Na^+ binding

The αM10 helix ends with three arginines (1003–1005) followed by the PGG motif and an extension of eight residues relative to the C terminus of the Ca^{2+}-ATPase (SERCA1a isoform). The small α-helix formed by the first part of this extension is accommodated between βM, αM7 and αM10, and the two C-terminal tyrosine residues are recognized by a binding pocket between αM7, αM8 and αM5 (Figs 5b and 6a, b). The insertion of Tyr 1015 and Tyr 1016 in this pocket is made possible by the kink of αM7 at Gly 848. Tyr 1016 seems to interact with Lys 766 of αM5 and Arg 933 in the loop connecting αM8 and αM9. This loop also contains Ser 936, a controversial phosphorylation site proposed to be responsible for some of the cAMP-dependent kinase (PKA)-mediated effects on the Na^+,K^+-ATPase[9,38]. Ser 936 is located within interaction distance of Arg 1003 (Fig. 6b). The unexpected features of the C terminus prompted us to study its functional importance by deletion of the five most C-terminal residues (Fig. 6c). The truncated enzyme (ΔKETYY) exhibited an extraordinary 26-fold reduction of the Na^+ affinity, yet the affinity for activating K^+ was like wild-type (Fig. 6c, upper panels). This is a direct effect of the truncation on the Na^+-binding E1 conformation, and not caused by displacement of the E1–E2 conformational equilibrium toward E2. In fact, the conformational equilibrium of ΔKETYY seems to be slightly displaced in the opposite direction towards E1, because the apparent affinities for ATP (binding preferentially to E1) and vanadate (binding only to E2) were found slightly enhanced and reduced, respectively (Fig. 6c, lower panels). The conspicuous and highly Na^+-selective effect of the ΔKETYY truncation is reminiscent of the effects observed previously for mutation of Tyr 771 (αM5)[39] and Thr 807 (αM6)[18]. Together with Glu 954 (αM9) these residues have been suggested to make up a third Na^+-binding site (Na^+ sites 1 and 2 probably being formed by almost the same coordinating side chains as the two K^+/Rb^+ binding sites)[40–42]. We find these residues to cluster and to be lined by Asp 808 (αM6), bridging to K^+/Rb^+ site 2. In addition, Gln 923 (αM8) is found in the same cluster and could be involved with the

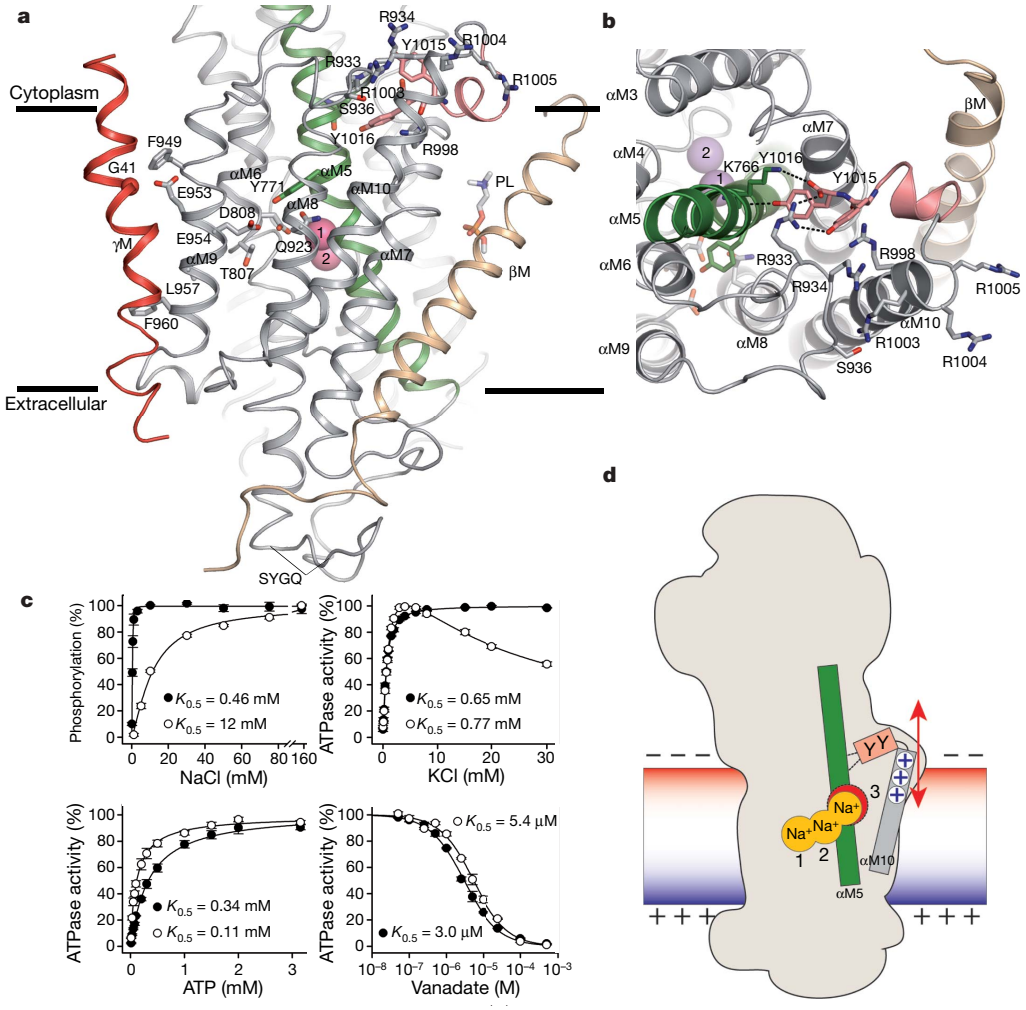


Figure 6 | The C-terminal switch. **a**, Side view of the transmembrane domain with the C-terminal switch region shown in the upper right part. In addition, residues of possible importance for interaction between α- and γ-subunits and between α- and β-subunits (SYGQ) and the third Na^+ site, as well as the modelled phosphatidylcholine head group (PL), are indicated. **b**, Top view from the cytoplasmic side of the C-terminal intrusion into the transmembrane region. Possible direct contacts between the tyrosine residues and Lys 766 and Arg 933 are represented by dashed lines. **c**, Functional analysis of the truncated enzyme ΔKETYY (open symbols) and the wild type (closed symbols). The apparent affinities for Na^+, K^+, ATP and vanadate are indicated as $K_{0.5}$ (the concentration giving half-maximum effect) values in the relevant panels. Error bars, s.e.m.; $n = 3$–6). The inhibition seen at high K^+ concentrations for ΔKETYY but not for the wild type, in the upper right panel, is a consequence of the reduced Na^+ affinity, allowing K^+ to compete efficiently with Na^+ at the sites of the E1 form[1]. **d**, Cartoon of the proposed functional elements of the C-terminal switch. The red double arrow indicates a change in membrane potential. The pull/push exerted by the switch on M5 may affect the affinity of the third electrogenic Na^+ site. The positive charges of the three arginines of αM10 at the membrane surface suggested to sense the membrane potential are indicated in blue. The interaction between the C-terminal tyrosine residue and M5 is indicated by lines.

third Na^+ site in the E1 form[42] (Fig. 6a and Supplementary Fig. 1). We propose that the direct contact of the C-terminal tail with αM5 and the loop between αM8 and αM9 serves to optimize Na^+ binding at the third site.

In light of the sensitivity of the Na^+,K^+-pump activity to the membrane potential[4,5], it is notable that Arg 1003, Arg 1004 and Arg 1005 at the end of αM10, together with Arg 933, Arg 934 and Arg 998, make the area around the C terminus in the membrane edge region highly electropositive (Fig. 6a, b). In various types of voltage-dependent ion channels arginine clusters act as voltage sensors that move in response to membrane depolarization[43,44], and in the Na^+,K^+-ATPase the arginine cluster associated with the C terminus could function similarly as a control point for a voltage-sensitive switch that alters the relations of the C terminus in its binding pocket during depolarization/repolarization, with consequences for the Na^+ affinity (Fig. 6d). The proposal of a direct structural and functional relation between the C terminus and the third Na^+ site is in accordance with the high voltage-sensitivity of the binding and release of one of the three Na^+ ions[4,5]. Interestingly, the human α1–α4 isoforms show a compelling pattern of differentiation in the 1003–1005 region (Supplementary Fig. 4), which may contribute to defining the differential sensitivity of the isoforms to variation in the membrane potential[45].

Conclusion

The present results provide clear structural evidence for the existence of a state in which the two counter-transported Rb^+/K^+ ions are occluded, as originally proposed on the basis of kinetic measurements[2,3]. The structural resemblance of the Na^+,K^+-ATPase α-subunit to the Ca^{2+}-ATPase is surprisingly high, even in the cation-binding pocket, thus raising the fundamental issue of how the specific cation selectivity is determined? Our results define a canonical set of cation-binding residues with only two conservative amino acid differences between the Na^+,K^+-ATPase and the Ca^{2+}-ATPase. We find it likely that subtle differences in the positions of the side chains and water molecules also contribute to define the cation selectivity. A unique aspect of the Na^+,K^+-ATPase is the non-canonical third Na^+ site, the location of which is hinted at by the present observations, even though our structure is the Rb^+/K^+ occluded enzyme. The C terminus of the Na^+,K^+-ATPase α-subunit has a previously unknown strategic location, allowing it to affect Na^+ binding and participate in Na^+,K^+-ATPase regulation.

METHODS SUMMARY

Na^+,K^+-ATPase was isolated from pig kidney outer medulla and purified by mild SDS treatment followed by isopycnic zonal centrifugation[46]. This preparation consists of α1- and β1-subunits together with the γ-subunit (γ_A and γ_B, Supplementary Fig. 5a, b). Crystals were obtained in the presence of 5 mM Rb^+ by the vapour-diffusion method in hanging drops (Supplementary Fig. 5c and Methods). The structure was determined on the basis of experimental electron-density maps. A low-resolution molecular replacement solution allowed site identification in derivative crystals for heavy-atom-based phasing. Phase extension by density modification and intercrystal averaging produced final experimental maps at 3.5 Å resolution, forming the basis for model building (Fig. 2d). The final model yields an *R*-factor of 27.7% and a free *R*-factor of 31.2% (Table 1). The magnesium fluoride complex at the catalytic site was modelled as the tetrahedral MgF_4^{2-}, as in the corresponding conformation of SERCA determined at 2.3 Å resolution[12].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 12 August; accepted 26 October 2007.

1. Skou, J. C. The influence of some cations on an adenosine triphosphatase from peripheral nerves. *Biochim. Biophys. Acta* **1000,** 439–446 (1957).
2. Post, R. L., Hegyvary, C. & Kume, S. Activation by adenosine triphosphate in the phosphorylation kinetics of sodium and potassium ion transport adenosine triphosphatase. *J. Biol. Chem.* **247,** 6530–6540 (1972).
3. Glynn, I. M. Annual review prize lecture. 'All hands to the sodium pump'. *J. Physiol. (Lond.)* **462,** 1–30 (1993).
4. Gadsby, D. C., Rakowski, R. F. & De Weer, P. Extracellular access to the Na,K Pump: pathway similar to ion channel. *Science* **260,** 100–103 (1993).
5. Apell, H. J. & Karlish, S. J. Functional properties of Na,K-ATPase, and their structural implications, as detected with biophysical techniques. *J. Membr. Biol.* **180,** 1–9 (2001).
6. Geering, K. The functional role of β subunits in oligomeric P-type ATPases. *J. Bioenerg. Biomembr.* **33,** 425–438 (2001).
7. Lutsenko, S. & Kaplan, J. H. An essential role for the extracellular domain of the Na,K-ATPase β-subunit in cation occlusion. *Biochemistry* **32,** 6737–6743 (1993).
8. Garty, H. & Karlish, S. J. Role of FXYD proteins in ion transport. *Annu. Rev. Physiol.* **68,** 431–459 (2006).
9. Therien, A. G. & Blostein, R. Mechanisms of sodium pump regulation. *Am. J. Physiol. Cell Physiol.* **279,** C541–C566 (2000).
10. Toyoshima, C., Nakasako, M., Nomura, H. & Ogawa, H. Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 A resolution. *Nature* **405,** 647–655 (2000).
11. Sorensen, T. L., Moller, J. V. & Nissen, P. Phosphoryl transfer and calcium ion occlusion in the calcium pump. *Science* **304,** 1672–1675 (2004).
12. Toyoshima, C., Nomura, H. & Tsuda, T. Lumenal gating mechanism revealed in calcium pump crystal structures with phosphate analogues. *Nature* **432,** 361–368 (2004).
13. Olesen, C., Sorensen, T. L., Nielsen, R. C., Moller, J. V. & Nissen, P. Dephosphorylation of the calcium pump coupled to counterion occlusion. *Science* **306,** 2251–2255 (2004).
14. Vilsen, B., Andersen, J. P., Petersen, J. & Jørgensen, P. L. Occlusion of $^{22}Na^+$ and $^{86}Rb^+$ in membrane-bound and soluble protomeric αβ-units of Na,K-ATPase. *J. Biol. Chem.* **262,** 10511–10517 (1987).
15. Danko, S., Yamasaki, K., Daiho, T. & Suzuki, H. Distinct natures of beryllium fluoride-bound, aluminum fluoride-bound, and magnesium fluoride-bound stable analogues of an ADP-insensitive phosphoenzyme intermediate of sarcoplasmic reticulum Ca^{2+}-ATPase: changes in catalytic and transport sites during phosphoenzyme hydrolysis. *J. Biol. Chem.* **279,** 14991–14998 (2004).
16. Vilsen, B. & Andersen, J. P. Mutation to the glutamate in the fourth membrane segment of Na^+,K^+- ATPase and Ca^{2+}-ATPase affects cation binding from both sides of the membrane and destabilizes the occluded enzyme forms. *Biochemistry* **37,** 10961–10971 (1998).
17. Einholm, A. P., Andersen, J. P. & Vilsen, B. Importance of Leu^{99} in transmembrane segment M1 of the Na^+,K^+-ATPase in the binding and occlusion of K^+. *J. Biol. Chem.* **282,** 23854–23866 (2007).
18. Vilsen, B. Mutant Glu781Ala of the rat kidney Na^+,K^+-ATPase displays low cation affinity and catalyzes ATP hydrolysis at a high rate in the absence of potassium ions. *Biochemistry* **34,** 1455–1463 (1995).
19. Kuntzweiler, T. A., Arguello, J. M. & Lingrel, J. B. Asp^{804} and Asp^{808} in the transmembrane domain of the Na,K-ATPase α subunit are cation coordinating residues. *J. Biol. Chem.* **271,** 29682–29687 (1996).
20. Blostein, R., Wilczynska, A., Karlish, S. J., Arguello, J. M. & Lingrel, J. B. Evidence that Ser^{775} in the α subunit of the Na,K-ATPase is a residue in the cation binding pocket. *J. Biol. Chem.* **272,** 24987–24993 (1997).
21. Pedersen, P. A., Nielsen, J. M., Rasmussen, J. H. & Jorgensen, P. L. Contribution to Tl^+, K^+, and Na^+ binding of Asn^{776}, Ser^{775}, Thr^{774}, Thr^{772}, and Tyr^{771} in cytoplasmic part of fifth transmembrane segment in α-subunit of renal Na,K-ATPase. *Biochemistry* **37,** 17818–17827 (1998).
22. Olesen, C. *et al.* The structural basis of calcium transport by the calcium pump. *Nature* doi:10.1038/nature06418 (this issue).
23. Clarke, D. M., Loo, T. W., Inesi, G. & MacLennan, D. H. Location of high affinity Ca^{2+}-binding sites within the predicted transmembrane domain of the sarcoplasmic reticulum Ca^{2+}-ATPase. *Nature* **339,** 476–478 (1989).
24. Andersen, J. P. & Vilsen, B. Amino acids Asn^{796} and Thr^{799} of the Ca^{2+}-ATPase of sarcoplasmic reticulum bind Ca^{2+} at different sites. *J. Biol. Chem.* **269,** 15931–15936 (1994).
25. Jardetzky, O. Simple allosteric model for membrane pumps. *Nature* **211,** 969–970 (1966).
26. Donnet, C., Arystarkhova, E. & Sweadner, K. J. Thermal denaturation of the Na,K-ATPase provides evidence for α-α oligomeric interaction and γ subunit association with the C-terminal domain. *J. Biol. Chem.* **276,** 7357–7365 (2001).
27. Colonna, T. E., Huynh, L. & Fambrough, D. M. Subunit interactions in the Na,K-ATPase explored with the yeast two-hybrid system. *J. Biol. Chem.* **272,** 12366–12372 (1997).
28. Rice, W. J., Young, H. S., Martin, D. W., Sachs, J. R. & Stokes, D. L. Structure of the Na^+,K^+-ATPase at 11-Å resolution: comparison with Ca^{2+}-ATPase in E1 and E2 states. *Biophys. J.* **80,** 2187–2197 (2001).
29. Hebert, H., Purhonen, P., Vorum, H., Thomsen, K. & Maunsbach, A. B. Three-dimensional structure of renal Na,K-ATPase from cryo-electron microscopy of two-dimensional crystals. *J. Mol. Biol.* **314,** 478–494 (2001).
30. Murata, T., Yamato, I., Kakinuma, Y., Leslie, A. G. & Walker, J. E. Structure of the rotor of the V-Type Na^+-ATPase from *Enterococcus hirae*. *Science* **308,** 654–659 (2005).
31. Teriete, P., Franzin, C. M., Choi, J. & Marassi, F. M. Structure of the Na,K-ATPase regulatory protein FXYD1 in micelles. *Biochemistry* **46,** 6774–6783 (2007).
32. Li, C. *et al.* Structural and functional interactions sites between Na^+,K^+-ATPase and FXYD proteins. *J. Biol. Chem.* **279,** 38895–38902 (2004).

33. Meij, I. C. *et al.* Dominant isolated renal magnesium loss is caused by misrouting of the Na^+,K^+-ATPase γ-subunit. *Nature Genet.* **26,** 265–266 (2000).
34. Einholm, A. P., Toustrup-Jensen, M., Andersen, J. P. & Vilsen, B. Mutation of Gly-94 in transmembrane segment M1 of Na^+,K^+-ATPase interferes with Na^+ and K^+ binding in E2P conformation. *Proc. Natl Acad. Sci. USA* **102,** 11254–11259 (2005).
35. Pedersen, B. P., Buch-Pedersen, M. J., Morth, J. P., Palmgren, M. G. & Nissen, P. Crystal structure of a plasma membrane proton pump. doi:10/1038/nature06417 *Nature* (this issue).
36. Håkansson, K. O. The crystallographic structure of Na,K-ATPase N-domain at 2.6 Å resolution. *J. Mol. Biol.* **332,** 1175–1182 (2003).
37. Hilge, M. *et al.* ATP-induced conformational changes of the nucleotide-binding domain of Na,K-ATPase. *Nature Struct. Biol.* **10,** 468–474 (2003).
38. Sweadner, K. J. & Feschenko, M. S. Predicted location and limited accessibility of protein kinase A phosphorylation site on Na-K-ATPase. *Am. J. Physiol. Cell Physiol.* **280,** C1017–C1026 (2001).
39. Vilsen, B., Ramlov, D. & Andersen, J. P. Functional consequences of mutations in the transmembrane core region for cation translocation and energy transduction in the Na^+,K^+-ATPase and the SR Ca^{2+}-ATPase. *Ann. NY Acad. Sci.* **834,** 297–309 (1997).
40. Ogawa, H. & Toyoshima, C. Homology modeling of the cation binding sites of Na^+K^+-ATPase. *Proc. Natl Acad. Sci. USA* **99,** 15977–15982 (2002).
41. Li, C., Capendeguy, O., Geering, K. & Horisberger, J. D. A third Na^+-binding site in the sodium pump. *Proc. Natl Acad. Sci. USA* **102,** 12706–12711 (2005).
42. Imagawa, T., Yamamoto, T., Kaya, S., Sakaguchi, K. & Taniguchi, K. Thr-774 (transmembrane segment M5), Val-920 (M8), and Glu-954 (M9) are involved in Na^+ transport, and Gln-923 (M8) is essential for Na,K-ATPase activity. *J. Biol. Chem.* **280,** 18736–18744 (2005).
43. Jiang, Y., Ruta, V., Chen, J., Lee, A. & MacKinnon, R. The principle of gating charge movement in a voltage-dependent K^+ channel. *Nature* **423,** 42–48 (2003).
44. Bass, R. B., Strop, P., Barclay, M. & Rees, D. C. Crystal structure of *Escherichia coli* MscS, a voltage-modulated and mechanosensitive channel. *Science* **298,** 1582–1587 (2002).
45. Crambert, G. *et al.* Transport and pharmacological properties of nine different human Na, K-ATPase isozymes. *J. Biol. Chem.* **275,** 1976–1986 (2000).
46. Jørgensen, P. L. Purification and characterization of (Na^+ + K^+)-ATPase. III. Purification from the outer medulla of mammalian kidney after selective removal of membrane components by SDS. *Biochim. Biophys. Acta* **356,** 36–52 (1974).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank E. Pohl, C. Schulze-Briese, and T. Tomizaki for extensive assistance with synchrotron data collection and Anna Marie Nielsen, Jytte Jørgensen, Kirsten Lykke Pedersen, and Lene Jacobsen for technical assistance. We are thankful to Hanne Poulsen, Anja P. Einholm, and Laure Yatime for discussion, and to Dr. Otto Hansen, University of Aarhus, for the γ-subunit specific polyclonal antibody. This work was supported financially through the DANSYNC programme, the Centre for Structural Biology of the Danish Research Council for Natural Sciences, and a Hallas-Møller stipend from the Novo-Nordisk Foundation (to P.N.), as well as by research grants from the Lundbeck Foundation, the Novo Nordisk Foundation (Fabrikant Vilhelm Pedersen og Hustrus Legat), and the Danish Medical Research Council (to B.V.). B.P.P. is supported by a fellowship from the Aarhus Graduate School of Science.

Author Contributions J.P.M. performed crystallization experiments, data collection and processing, structure determination and refinement, and structural analysis. B.P.P. assisted in data collection, structure determination and analysis. T.L.S. identified initial crystallization conditions. M.S.T.J. produced mutant enzyme and characterized it functionally. J.P. performed occlusion experiments and prepared and characterized Na^+,K^+-ATPase samples as required for crystallization. P.N., B.V., and J.P.A. supervised the project and analysed the structure. The paper was written by J.P.M., P.N., B.V., and J.P.A., P.N. and B.V. have contributed equally and with equal resources to the project.

Author Information The Na^+,K^+-ATPase has the PDB code 3B8E. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to P.N. (pn@mb.au.dk) and B.V. (bv@fi.au.dk).

METHODS

Enzyme preparation and biochemical studies. Following zonal centrifugation[46] membrane fractions with particularly high specific Na^+,K^+-ATPase activity were selected for crystallization experiments. The membranes were incubated in 5 mM RbF (or 20 mM KCl and 5 mM KF for crystallization of K^+-bound enzyme), 5 mM $MgCl_2$, 100 mM *N*-methyl-D-glucamine (NMDG), 40 mM MOPS, pH 7.0. For solubilization of the enzyme, the non-ionic detergent octaethyleneglycol mono-*n*-dodecylether ($C_{12}E_8$)[14] was added at a ratio of 1.12 mg $C_{12}E_8$ per mg membrane protein. Before crystallization, the insoluble material (20–25% of total protein) was removed by ultracentrifugation.

$^{86}Rb^+$ occlusion in the membranous and soluble enzyme preparations was measured according to the previously described principles[14]. Enzyme preincubated with 0.1 mM $^{86}Rb^+$ in the presence or absence of magnesium fluoride (2–5 mM) was mixed with a dissociation solution containing an 83-fold excess of non-radioactive Rb^+ with and without 3 mM ATP at 25 °C. The amount of $^{86}Rb^+$ remaining bound to the enzyme was determined at the indicated time intervals by subjecting aliquots of the samples to rapid ionic exchange chromatography, following cooling to 2 °C.

Deletion of the five most C-terminal residues KETYY was carried out by PCR of complementary DNA encoding the rat α1 isoform of the Na^+,K^+-ATPase followed by expression in COS cells, and the previously described assays for phosphorylation and ATPase activity[18,47] were used for the functional characterization (Fig. 6c). The Na^+ dependence of phosphorylation (Fig. 6c, upper left panel) was determined in the presence of 2 μM [γ-^{32}P]ATP in the absence of K^+. The K^+ dependence of Na^+,K^+-ATPase activity (Fig. 6c, upper right panel) was determined in the presence of 40 mM Na^+ and 3 mM ATP. The ATP and vanadate dependencies of the Na^+,K^+-ATPase activity (Fig. 6c, lower left and right panels, respectively) were determined in the presence of 130 mM Na^+ and 20 mM K^+. The ΔKETYY mutant exhibited a maximal catalytic turnover rate of $9720 \pm 490\ min^{-1}$ (mean ± s.e.m., $n = 6$) versus $8470 \pm 170\ min^{-1}$ ($n = 11$) for the wild type, determined at 37 °C, pH 7.4, in the presence of 3 mM MgATP and saturating Na^+ and K^+ concentrations of 130 mM and 20 mM K^+, respectively.

Crystallization. Crystals were grown by vapour diffusion from hanging drop at 19 °C. Protein solution was mixed with precipitating solution (14% PEG2000mme, 200 mM choline chloride, 4 mM DTT, 4% glycerol, 4% MPD) in a 1:1 ratio by adding 4 μl protein, 4 μl precipitating solution, and 0.8 μl 0.1–0.35% β-DDM. The initial precipitate formed was spun down before 2 μl hanging drops were dispensed. The very thin and fragile crystals (Supplementary Fig. 5c) appeared after 3–4 days and grew to their maximum size ($0.6 \times 0.2 \times 0.05\ mm^3$) within a month. The crystals were mounted in Litholoops (Molecular Dimensions) from the mother liquor. Before flash-cooling in liquid nitrogen excess mother liquor was dipped away by gently touching a glass cover slip with the edge of the loop. For heavy-atom derivatization, dry powder of $Ta_6Br_{12}^{2+}$ or Orange-Pt was dusted directly to the drop until the crystals appeared light green or faint orange, respectively.

Structure determination and analysis. All data sets were collected at 100 K on the end stations X06SA and X10SA at the Swiss Light Source (SLS) in Villingen. The diffraction data were processed and scaled with XDS[48]. The crystal form exhibits $P2_12_12_1$ space-group symmetry with unit cell dimensions $a = 68.93$ Å, $b = 261.5$ Å and $c = 333.8$ Å. The asymmetric unit contains 2 αβγ complexes. Initial phases were obtained by molecular replacement at 6 Å resolution using the program PHASER[49] and a search model derived from Ca^{2+}-ATPase in the E2-AlF_4 form (PDB code 1XP5)[13]. Heavy-atom sites were then identified by difference Fourier maps using the molecular replacement phases and MIRAS (multiple isomorphous replacement with anomalous scattering) phases obtained at 6 Å resolution with SHARP/autoSHARP 2.0 (ref. 50). The MIRAS phases were refined and further extended using DMMULTI[51] and RESOLVE[52] to 3.5 Å resolution, exploiting solvent flattening (75% solvent), two-fold non-crystallographic symmetry-averaging and two-fold inter-crystal averaging, using two data sets exhibiting a low degree of isomorphism. The final experimental map was of sufficient quality to trace the entire model. Model building was performed using O[53] and model refinement was performed with CNS1.2 (ref. 54). For Fig. 3c, a data set obtained for Na^+,K^+-ATPase with bound K^+ was used to perform a simulated annealing omit map, calculated at 4 Å resolution, using model phases, omitting both Rb^+ ions and residues in a radius of 4.5 Å. All structural representations in this paper were prepared with Pymol (http://www.pymol.org).

47. Rodacker, V., Toustrup-Jensen, M. & Vilsen, B. Mutations Phe^{785}Leu and Thr^{618}Met in Na^+,K^+-ATPase, associated with familial rapid-onset dystonia parkinsonism, interfere with Na^+ interaction by distinct mechanisms. *J. Biol. Chem.* **281,** 18539–18548 (2006).
48. Kabsch, W. Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J. Appl. Cryst.* **26,** 795–800 (1993).
49. McCoy, A. J., Grosse-Kunstleve, R. W., Storoni, L. C. & Read, R. J. Likelihood-enhanced fast translation functions. *Acta Crystallogr. D* **61,** 458–464 (2005).
50. Vonrhein, C., Blanc, E., Roversi, P. & Bricogne, G. Automated structure solution with autoSHARP. *Methods Mol. Biol.* **364,** 215–230 (2006).
51. Cowtan, K. D. & Main, P. Phase combination and cross validation in iterated density-modification calculations. *Acta Crystallogr. D* **52,** 43–48 (1996).
52. Terwilliger, T. C. Maximum-likelihood density modification. *Acta Crystallogr. D* **56,** 965–972 (2000).
53. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47,** 110–119 (1991).
54. Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54,** 905–921 (1998).

Vol 450|13 December 2007|doi:10.1038/nature06425

LETTERS

A dynamic, rotating ring current around Saturn

S. M. Krimigis[1,2], N. Sergis[2], D. G. Mitchell[1], D. C. Hamilton[3] & N. Krupp[4]

The concept of an electrical current encircling the Earth at high altitudes was first proposed in 1917 to explain the depression of the horizontal component of the Earth's magnetic field during geomagnetic storms[1–4]. *In situ* measurements of the extent and composition of this current were made some 50 years later[5] and an image was obtained in 2001 (ref. 6). Ring currents of a different nature were observed at Jupiter[7,8] and their presence inferred at Saturn[9,10]. Here we report images of the ring current at Saturn, together with a day–night pressure asymmetry and tilt of the planet's plasma sheet, based on measurements using the magnetospheric imaging instrument (MIMI) on board Cassini. The ring current can be highly variable with strong longitudinal asymmetries that corotate nearly rigidly with the planet. This contrasts with the Earth's ring current, where there is no rotational modulation and initial asymmetries are organized by local time effects.

The MIMI instrument[11] on the Cassini Saturn orbiter spacecraft comprises three sensors that measure particles in specific energy ranges: (1) the ion and neutral camera (INCA) obtains images of ion and neutral species (~3 to >200 keV per nucleon); (2) the charge energy mass spectrometer (CHEMS) measures ions and their charge states (~3 to 230 keV per charge); and (3) the low-energy magnetospheric measurement system (LEMMS) measures ions (~0.02 to ~18 MeV) and electrons (~0.015 to ~1 MeV). Since its orbit insertion at Saturn on 1 July 2004, Cassini has been making nearly continuous measurements of the charged particle environment. The initial set of orbits were near Saturn's equatorial plane and provided *in situ* sampling of both the plasma sheet[12] and the ring current[13]. It was found that the ring current consists of a high (>~1) beta (ratio of particle to magnetic pressure) region extending from $\sim 10 < L < 19 R_S$ (where one Saturn radius $R_S = 60{,}268$ km and L is the magnetic shell parameter), with most of the pressure residing in the range $10 < \sim E < \sim 150$ keV and O^+ ions generally contributing >50% of the total. Lower-energy (<3 keV) plasma was not an important contributor to the pressure[13,14] and is not included in the present study.

Beginning in autumn 2006, the inclination of Cassini's orbit was raised gradually to high latitudes, providing the opportunity to make off-equatorial measurements *in situ* with the CHEMS and LEMMS sensors and to obtain images of the ion distribution around the equator using the energetic neutral atoms (ENA) technique by looking down (or up) with the INCA sensor. Briefly, the ENA technique relies on charge exchange between trapped ions and a residual neutral gas that results in fast atoms escaping the system and being sensed as if they were photons. One such image is shown in Fig. 1. The ring current maximum intensity is generally outside the orbit of Rhea; observable intensities may extend beyond the orbit of Titan. Overall, the image in Fig. 1 illustrates that although this interval was chosen specifically as an example with minimal local time/longitudinal structure, the ring current, unsurprisingly, is not the uniform, symmetric construct postulated in early modelling of Saturn's magnetic field[9,15].

Figure 2 shows *in situ* measurements of ion pressure and time-intensity profiles from the CHEMS sensor for days 26–44 of 2007. The pressure profile (Fig. 2a) displays a strong asymmetry in local time, with the night side plasma sheet being much narrower in latitudinal extent (~30° versus ~95° on the day side) and less intense by a factor of ~10. Both are centred about Saturn's equatorial plane at a distance of ~20 R_S as seen from the coordinates (Fig. 2b–d). The particle intensity spectrogram (Fig. 2e) from CHEMS shows that the energy spectrum is somewhat softer on the day side, and it extends to the upper energy limit (230 keV per charge) of the sensor. Detailed examination of other MIMI data (not shown here) shows that Cassini remained within the magnetosphere at all times. The drop-off in particle pressure early on day 38 is most probably caused by crossing from the plasma sheet into the lobe, perhaps past the last closed field line at ~44° latitude at local time ~16:15. Note that the equivalent crossing on the night side occurred at ~17° latitude, somewhat closer to the planet (~24 R_S versus ~26 R_S).

To investigate this clear asymmetry further, we have performed an analysis of all Cassini orbits from 1 July 2004 (Saturn orbit insertion)

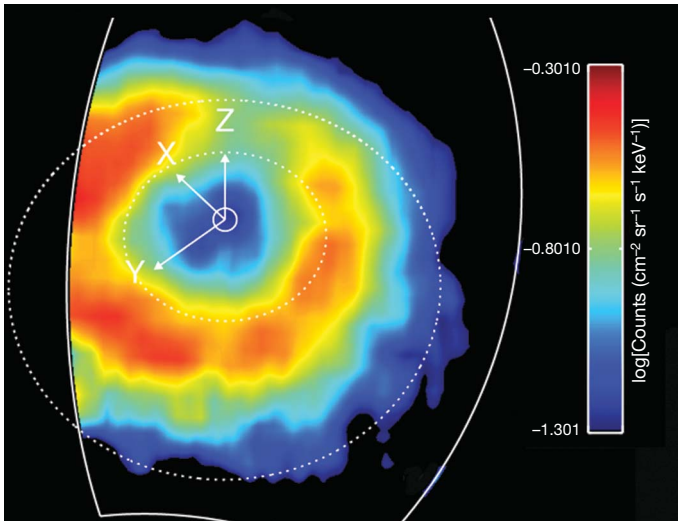


Figure 1 | ENA image of the ring current as viewed from above the northern hemisphere. This image, in the range 20–50 keV, was obtained on 19 March 2007, with MIMI/INCA, at a latitude of ~54.5° and radial distance ~24.5R_S. Saturn is at the centre, and the dotted circles represent the orbits of Rhea (8.74R_S) and Titan (20.2R_S). The Z axis points parallel to Saturn's spin axis, the X axis points roughly sunward in the Sun–spin-axis plane, and the Y axis completes the system, pointing roughly towards dusk. The INCA field of view is marked by the white line and accounts for the cut-off of the image on the left. The image is a co-spatial average of several frames over the period 16:32–19:44 UTC (Universal time).

[1]The Johns Hopkins University Applied Physics Laboratory, Laurel, Maryland 20723, USA. [2]Office of Space Research and Technology, Academy of Athens, Soranou Efesiou 4, 11527 Athens, Greece. [3]Department of Physics, University of Maryland, College Park, Maryland 20742, USA. [4]Max-Planck-Institut für Sonnensystemforschung, Max-Planck-Strasse 2, 37191 Katlenburg-Lindau, Germany.

to June 2007. In Fig. 3 the results are plotted in the ρ–Z plane and separated into the day side and night side parts. Figure 3a includes all measured off-equatorial values but excludes the dawn–dusk portion to obtain a clear separation of day–night effects. Although the orbital coverage in Z is not uniform, the higher pressures on the day side appear to extend to much higher latitudes than the night side, certainly at $<20R_S$. This is clearly evident in Fig. 3b, where pressures $<5\times10^{-11}$ dynes cm^{-2} have been omitted. Not only is the day–night asymmetry striking, but also the shape of the night side plasma sheet beyond $\sim20R_S$ is outlined and is seen to be tilted northward at an angle of $\sim10^{\circ}$, although the orbital coverage in this region is not extensive. A three-dimensional projection of the pressure distribution at each plane is shown in Supplementary Fig. 1. Examination of each Cassini orbit at all available local times suggests that the day side plasma sheet thins gradually towards the night side, even though the detailed distribution with local time is not fully determined because of incomplete latitudinal/local time coverage.

Our interpretation thus far, based on the pressure distribution, is shown in Fig. 4. This view from above Saturn's equatorial plane illustrates the compressed day side plasma sheet and indicates its expansion to northern and southern latitudes. We expect that the sheet gradually thins on the dusk side but is drawn tailward at midnight and inflates again at dawn. Whether there is loss of plasma on the night side is not clear, because this sketch represents an average picture of all orbits over a nearly three-year period. We have, however, repeatedly observed acceleration events both in the magnetotail[16] and in parts of the magnetosphere, where the injected plasma cloud clearly corotates with the planet.

One such event is shown in Fig. 5, a sequence of six INCA images covering a Saturn rotation. The top left panel shows a large, factor of ~10, intensity increase between dawn and local midnight that moves

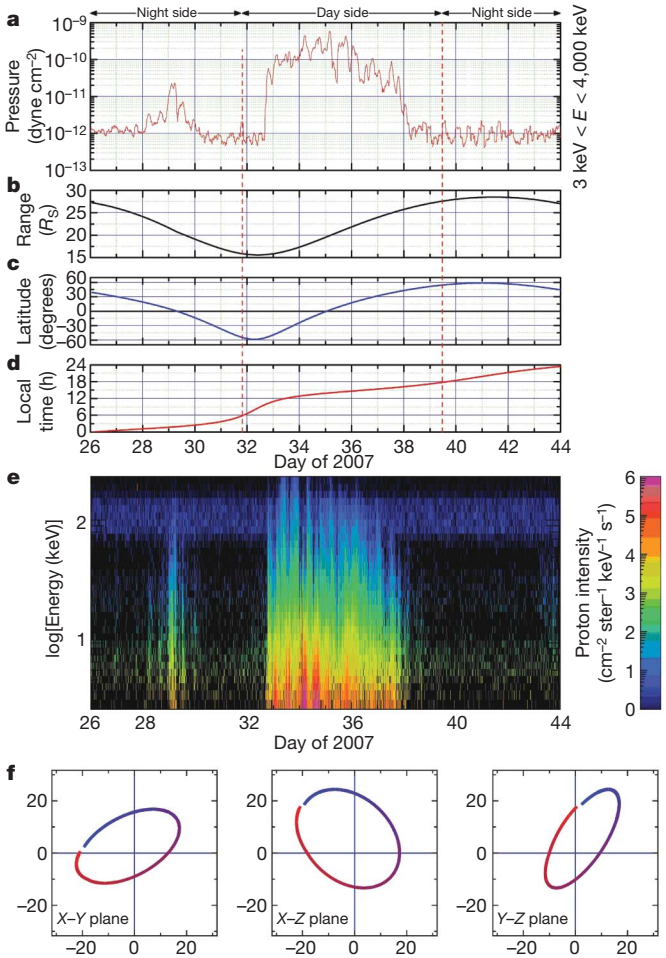


Figure 2 | Typical Cassini pass through the day–night plasma sheet in early 2007. **a**, Ion pressure profile over the indicated time interval. **b–d**, Radial distance, latitude, and local time coordinates of the Cassini spacecraft. **e**, Differential intensity–time spectrogram of protons from the CHEMS sensor over the energy interval $3 < E < 230$ keV per charge; the horizontal dark blue band near the top is background interference and is not relevant. **f**, Projection of the Cassini orbit in the X–Y, X–Z and Y–Z planes in Saturn-centred coordinates; the red and blue colours correspond to inbound and outbound parts of the orbit, respectively.

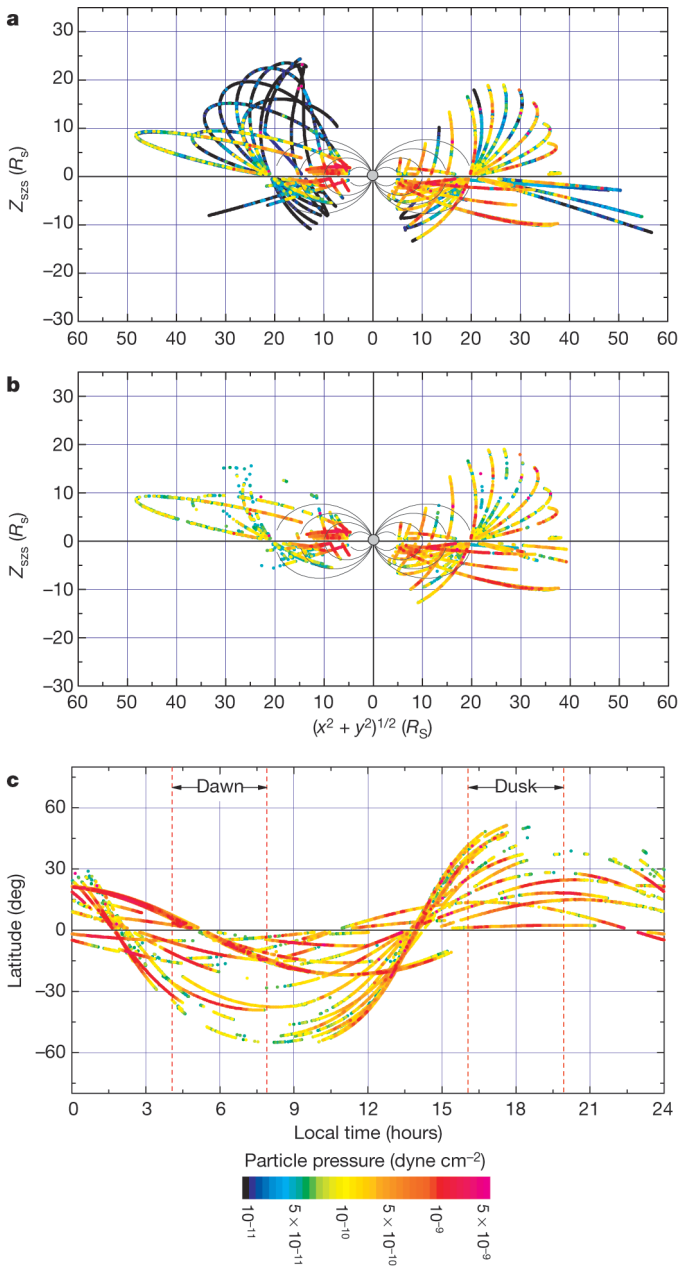


Figure 3 | Observed particle pressure profile (colour scale) for all non-equatorial Cassini orbits. The pressure profile (~3 to 4,000 keV) was computed over the period from July 2004 to June 2007 and is projected onto the $\rho(X^2 + Y^2)^{1/2}$–Z plane, where X and Y are positions on those axes. **a**, Pressure profile on the day side (08:00–16:00 h, right) and night side (20:00–04:00 h, left) over the full dynamic range measured by the CHEMS and LEMMS sensors (5×10^{-13} to 5×10^{-9} dynes cm^{-2}), clearly illustrating the orbital coverage. **b**, The same data as in **a** but for a threshold $>5\times10^{-11}$ dynes cm^{-2}; the day–night asymmetry at $R > 20R_S$ is striking. **c**, Pressure coverage in local time and along the Z axis for all thresholded data, but also including the dawn–dusk coverage not shown in **a** or **b**.

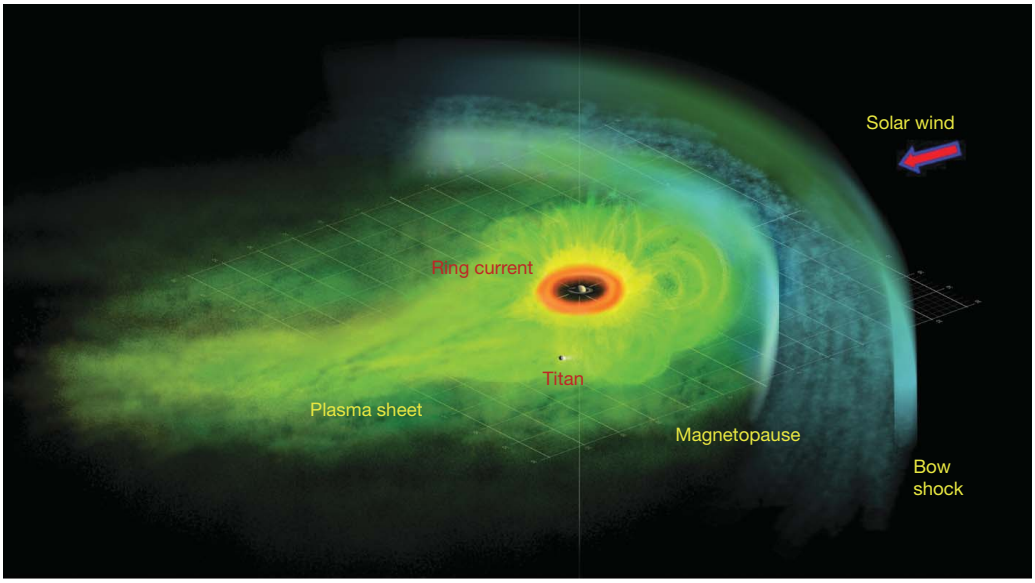


Figure 4 | An artist's concept of Saturn's plasma sheet and embedded ring current, consistent with the data shown in Fig. 3. Saturn is at the centre, with the red 'doughnut' representing the distribution of dense neutral gas (H, O, O_2 and OH) outside the rings. Beyond this region, energetic ions populate the plasma sheet to the day side magnetopause, filling the faintly sketched magnetic flux tubes to higher latitudes and contributing to the ring current. The plasma sheet thins gradually towards the night side. The view is from above Saturn's equatorial plane, which is represented by grid lines. Titan's location is shown for scale. The location of the bow shock is marked, as is the flow of the deflected solar wind in the magnetosheath.

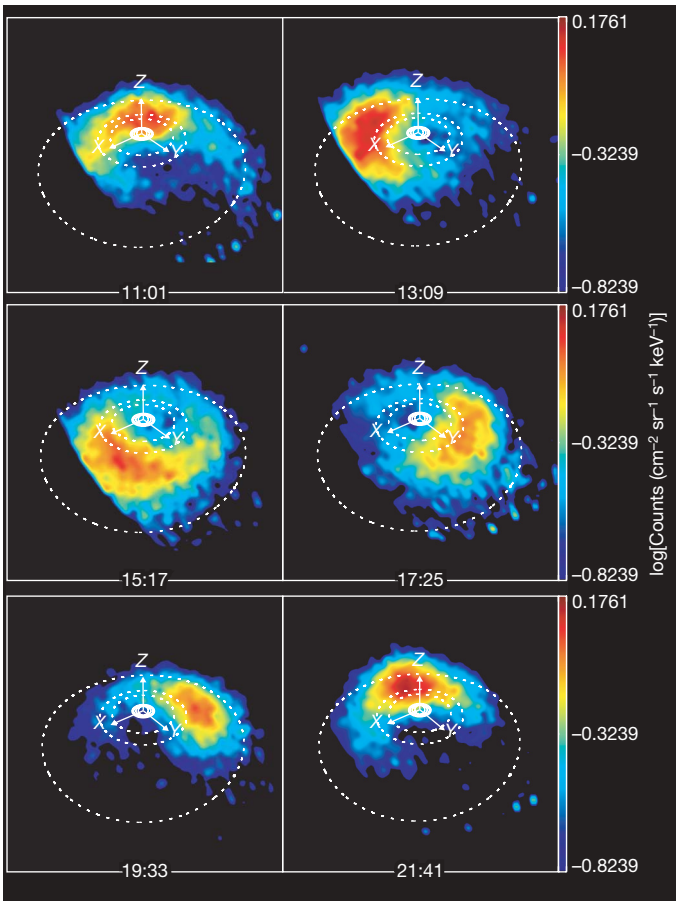


Figure 5 | Sequence of six ENA images in neutral hydrogen taken by INCA. This sequence of ENA images in the range 20–50 keV was obtained on 24 February 2007, and covered a full Saturn rotation. Cassini was located at ~32° latitude and 26R_S from Saturn at local time ~15:12. Saturn is at the centre, the *X* axis is pointing approximately in the solar direction, *Y* is pointing towards dusk, and *Z* is pointing along Saturn's spin vector. Dotted lines show the orbits of Dione (6.26R_S), Rhea (8.74R_S), and Titan (20.2R_S) in proper perspective. Sharp edges on the left of the first three frames are attributable to the limit of INCA's field of view. The images are spaced at roughly 2-h intervals. See the Supplementary Video.

anticlockwise through dawn, then day side, then local evening (middle right panel), then local midnight, and then returns to its original location some 11 h later (bottom right panel). Such sequences are seen quite often and are captured in full resolution as movie clips (Supplementary Video). It is clear that in this case, the increase is very nearly fixed in longitude, not local time, and the corresponding ring current is asymmetric and variable. The contribution of such events to the average plasma sheet and ring current sketched in Fig. 4 has not been assessed and is the subject of future study.

The results presented above, using both *in situ* and imaging measurements within Saturn's magnetosphere, demonstrate that the energetic particle pressure contribution to the ring current is far more complex than the ring current initially modelled[9] as a symmetric region extending from ~9R_S to ~15R_S in the equatorial plane with a thickness of ±2.5R_S. Although this early model has been useful until now, the Cassini/MIMI results show that it is a dynamic current system embedded in Saturn's plasma sheet from ~9R_S to ~20R_S, with a vertical extent as high as ±45° on the day side. The night side hot-plasma region (ring current/plasma sheet) is tilted above Saturn's equatorial plane with a vertical dimension of ±5R_S close in (<15R_S) but extending mostly above that plane to distances of ~50R_S and vertically to ~10R_S. This asymmetry in local time is somewhat consistent with expectations of magnetohydrodynamic simulations[17], and the northward deflection is expected from the tilt of the spin-aligned dipole with respect to the incoming solar wind. The 'hinge' point is predicted to be at ~20R_S, not inconsistent with the present observations. The day side inflation of the plasma sheet, however, is much greater than that modelled so far[17].

The apparent corotation with the planet of episodically injected plasma, as documented in Fig. 5, represents a conceptual conundrum, in that such plasma blobs are apparently associated with a particular Saturn longitude and are not organized by local time. How the steady-state plasma population can display such obvious day–night asymmetry while the injected particles corotate is a mystery that will require analyses of many more such events.

Received 9 August; accepted 18 October 2007.

1. Schmidt, A. Erdmagnetismus. In *Enzyklopädie der Mathematischen Wissenschaften, Band VI* (Teubner, Leipzig, 1917).
2. Chapman, S. An outline of the theory of magnetic storms. *Proc. R. Soc. Lond. A* **95**, 61–83 (1918).
3. Chapman, S. & Ferraro, V. C. A. A new theory of magnetic storms, Part I. The initial phase. *Terr. Magn. Atmos. Elect.* **36**, 171–186 (1931).
4. Chapman, S. & Ferraro, V. C. A. A new theory of magnetic storms. *Terr. Magn. Atmos. Elect.* **38**, 79–96 (1933).

5. Krimigis, S. M. *et al.* Magnetic storm of September 4, 1984: A synthesis of ring current spectra and energy densities measured with AMPTE/CCE. *Geophys. Res. Lett.* **12,** 329–332 (1995).
6. Mitchell, D. G. *et al.* Imaging two geomagnetic storms in energetic neutral atoms. *Geophys. Res. Lett.* **28,** 1151–1155 (2001).
7. Krimigis, S. M. *et al.* Characteristics of hot plasma in the Jovian magnetosphere: Results from the Voyager spacecraft. *J. Geophys. Res.* **86,** 8227–8257 (1981).
8. Mauk, B. M. *et al.* Energetic ion characteristics and neutral gas interactions in Jupiter's magnetosphere. *J. Geophys. Res.* **109,** A09512, doi:10.1029/2003JA010270 (2004).
9. Connerney, J. E. P., Acunã, M. & Ness, N. F. Saturn's ring current and inner magnetosphere. *Nature* **292,** 724–726 (1981).
10. Krimigis, S. M. *et al.* General characteristics of hot plasma and energetic particles in the Saturnian magnetosphere: Results from the Voyager spacecraft. *J. Geophys. Res.* **88,** 8871–8892 (1983).
11. Krimigis, S. M. *et al.* Magnetospheric imaging instrument (MIMI) on the Cassini mission to Saturn/Titan. *Space Sci. Rev.* **114,** 233–329 (2004).
12. Krupp, N. *et al.* The Saturnian plasma sheet as revealed by energetic particle measurements. *Geophys. Res. Lett.* **32,** L20503, doi:10.1029/2005GL022829 (2005).
13. Sergis, N. *et al.* Ring current at Saturn: Energetic particle pressure in Saturn's equatorial magnetosphere measured with Cassini/MIMI. *Geophys. Res. Lett.* **34,** L09102, doi:10.1029/2006GL029223 (2007).
14. Sittler, E. C. Jr *et al.* Cassini observations of Saturn's inner plasmasphere: Saturn orbit insertion results. *Planet. Space Sci.* **54,** 1197–1210 (2006).
15. Mauk, B. H., Krimigis, S. M. & Lepping, R. P. Particle and field stress balance within a planetary magnetosphere. *J. Geophys. Res.* **90,** 8253–8264 (1985).
16. Mitchell, D. G. *et al.* Energetic ion acceleration in Saturn's magnetotail: Substorms at Saturn? *Geophys. Res. Lett.* **32,** L20501, doi:10.1029/2005GL022647 (2005).
17. Hansen, K. C. *et al.* Global MHD simulations of Saturn's magnetosphere at the time of Cassini approach. *Geophys. Res. Lett.* **32,** L20506, doi:10.1029/2005GL022835 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Kusterer (The Johns Hopkins University Applied Physics Laboratory) for assistance with the data reduction. We are grateful to colleagues on the MIMI team who provided comments that have improved the presentation. Work at The Johns Hopkins University Applied Physics Laboratory was supported by NASA and by subcontracts at the University of Maryland and the Office of Space Research and Technology of the Academy of Athens. The German contribution of MIMI/LEMMS was financed in part by the Bundesministerium für Bildung und Forschung (BMBF) through the Deutsches Zentrum für Luft-und Raumfahrt e.V. (DLR) and by the Max-Planck-Gesellschaft.

Author Contributions S.M.K. is the MIMI Principal Investigator and contributed most of the text; N.S. analysed the *in situ* pressure data; D.G.M. is the INCA lead investigator and analysed the ENA images; D.C.H. is the lead investigator of CHEMS, while N.K. oversees the LEMMS data analyses.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to S.M.K. (tom.krimigis@jhuapl.edu).

LETTERS

Optical rogue waves

D. R. Solli[1], C. Ropers[1,2], P. Koonath[1] & B. Jalali[1]

Recent observations show that the probability of encountering an extremely large rogue wave in the open ocean is much larger than expected from ordinary wave-amplitude statistics[1–3]. Although considerable effort has been directed towards understanding the physics behind these mysterious and potentially destructive events, the complete picture remains uncertain. Furthermore, rogue waves have not yet been observed in other physical systems. Here, we introduce the concept of optical rogue waves, a counterpart of the infamous rare water waves. Using a new real-time detection technique, we study a system that exposes extremely steep, large waves as rare outcomes from an almost identically prepared initial population of waves. Specifically, we report the observation of rogue waves in an optical system, based on a microstructured optical fibre, near the threshold of soliton-fission supercontinuum generation[4,5]—a noise-sensitive[5–7] nonlinear process in which extremely broadband radiation is generated from a narrowband input[8]. We model the generation of these rogue waves using the generalized nonlinear Schrödinger equation[9] and demonstrate that they arise infrequently from initially smooth pulses owing to power transfer seeded by a small noise perturbation.

For centuries, seafarers have told tales of giant waves that can appear without warning on the high seas. These mountainous waves were said to be capable of destroying a vessel or swallowing it beneath the surface, and then disappearing without the slightest trace. Until recently, these tales were thought to be mythical. In the mid-1990s, however, freak waves proved very real when recorded for the first time by scientific measurements during an encounter at the Draupner oil platform in the North Sea[3]. Although they are elusive and intrinsically difficult to monitor because of their fleeting existences, satellite surveillance has confirmed that rogue waves roam the open oceans, occasionally encountering a ship or sea platform, sometimes with devastating results[1]. It is now believed that a number of infamous maritime disasters were caused by such encounters[10].

The unusual statistics of rogue waves represent one of their defining characteristics. Conventional models of ocean waves indicate that the probability of observing large waves should diminish extremely rapidly with wave height, suggesting that the likelihood of observing even a single freak wave in hundreds of years should be essentially non-existent. In reality, however, ocean waves appear to follow 'L-shaped' statistics: most waves have small amplitudes, but extreme outliers also occur much more frequently than expected in ordinary (for example, gaussian or Rayleigh) wave statistics[11–13].

It is likely that more than one process can produce occasional extreme waves with small but non-negligible probability[14,15]. Possible mechanisms that have been suggested to explain oceanic rogue waves include effects such as nonlinear focusing via modulation instability in one dimension[16,17] and in two-dimensional crossings[18,19], nonlinear spectral instability[20], focusing with caustic currents[21] and anomalous wind excitation[12]. Nonlinear mechanisms have attracted particular attention because they possess the requisite extreme sensitivity to initial conditions.

Although the physics behind rogue waves is still under investigation, observations indicate that they have unusually steep, solitary or tightly grouped profiles, which appear like ''walls of water''[10]. These features imply that rogue waves have relatively broadband frequency content compared with normal waves, and also suggest a possible connection with solitons—solitary waves, first observed by J. S. Russell in the nineteenth century, that propagate without spreading in water because of a balance between dispersion and nonlinearity. As rogue waves are exceedingly difficult to study directly, the relationship between rogue waves and solitons has not yet been definitively established, but it is believed that they are connected.

So far, the study of rogue waves in the scientific literature has focused on hydrodynamic studies and experiments. Intriguingly, there are other physical systems that possess similar nonlinear characteristics and may also support rogue waves. Here we report the observation and numerical modelling of optical rogue waves in a system based on probabilistic supercontinuum generation in a highly nonlinear microstructured optical fibre. We coin the term 'optical rogue waves' based on striking phenomenological and physical similarities between the extreme events of this optical system and oceanic rogue waves.

Supercontinuum generation has received a great deal of attention in recent years for its complex physics and wealth of potential applications[5,8]. An extremely broadband supercontinuum source can be created by launching intense seed pulses into a nonlinear fibre at or near its zero-dispersion wavelength[22]. In this situation, supercontinuum production involves generation of high-order solitons—the optical counterparts of Russell's solitary water waves—which fission into redshifted solitonic and blueshifted non-solitonic components at different frequencies[4,5]. The solitonic pulses shift further towards the red as they propagate through the nonlinear medium because of the Raman-induced self-frequency shift[9]. Interestingly, frequency downshifting effects are also known to occur in water wave propagation. It has been noted that the aforementioned Raman self-frequency shift represents an analogous effect in optics[23]. The nonlinear processes responsible for supercontinuum generation amplify the noise present in the initial laser pulse[6,7]. Especially for long pulses and continuous-wave input radiation, modulation instability—an incoherent nonlinear wave-mixing process—broadens the spectrum from seed noise in the initial stages of propagation and, as a result, the output spectrum is highly sensitive to the initial conditions[24,25].

A critical challenge in observing optical rogue waves is the lack of real-time instruments that can capture a large number of very short random events in a single shot. To solve this problem, we use a wavelength-to-time transformation technique inspired by the concept of photonic time-stretch analog-to-digital conversion[26]. In the present technique, group-velocity dispersion (GVD) is used to stretch the waves temporally so that many thousands of random ultra-short events can be captured in real time. A different single-shot technique has been used to study isolated supercontinuum pulses[27]; however, the real-time capture of a large number of random

[1]Department of Electrical Engineering, University of California, Los Angeles 90095, USA. [2]Max Born Institute for Nonlinear Optics and Short Pulse Spectroscopy, D-12489 Berlin, Germany.

events has not been reported. Using the present method, a small but statistically significant fraction of extreme waves can be discerned from a large number of ordinary events, permitting the first observation of optical rogue waves.

The supercontinuum radiation used in our experiments is generated by sending picosecond seed pulses at 1,064 nm through a length of highly nonlinear microstructured optical fibre with matched zero-dispersion wavelength. The output is red-pass filtered at 1,450 nm and stretched as described above so that many thousands of events can be captured with high resolution in a single-shot measurement. A schematic of the experimental apparatus is displayed in Fig. 1 and additional details are provided in the Methods Summary.

Using this setup, we acquire large sets of pulses in real time for very low seed pulse power levels—power levels below the threshold required to produce appreciable supercontinuum. We find that the pulse-height distributions are sharply peaked with a well-defined mean, but contrary to expectation, rare events with far greater intensities also appear. In Fig. 1, we show representative single-shot time traces and histograms for three different low power levels. In these traces, the vast majority of events are concentrated in a small number of bins and are so weak that they are buried beneath the noise floor of the measurement process; however, the most extreme ones reach intensities at least 30–40 times the average. The histograms display a clear L-shaped profile, with extreme events occurring rarely, yet

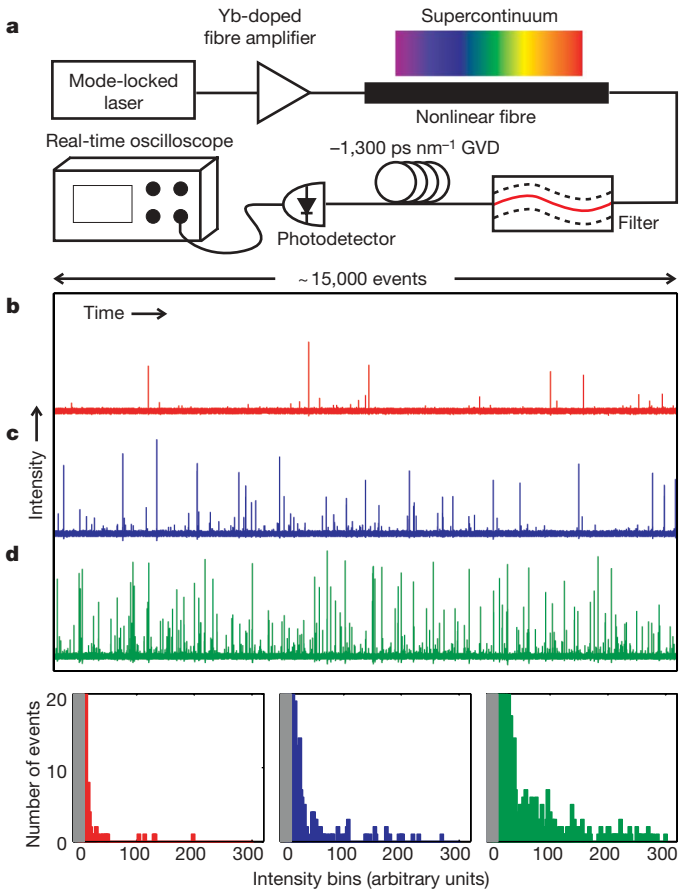


Figure 1 | Experimental observation of optical rogue waves. a, Schematic of experimental apparatus. **b–d**, Single-shot time traces containing roughly 15,000 pulses each and associated histograms (bottom of figure: left, **b**; middle, **c**; right, **d**) for average power levels 0.8 μW (red), 3.2 μW (blue) and 12.8 μW (green), respectively. The grey shaded area in each histogram demarcates the noise floor of the measurement process. In each measurement, the vast majority of events (>99.5% for the lowest power) are buried in this low intensity range, and the rogue events reach intensities of at least 30–40 times the average value. These distributions are very different from those encountered in most stochastic processes.

much more frequently than expected based on the relatively narrow distribution of typical events.

Because the red-pass filter transmits only a spectral region that is nearly dark in the vast majority of events, the rare events clearly have extremely broadband, frequency-downshifted spectral content. The data also show that the frequency of occurrence of the rogue events increases with the average power, but the maximum height of a freak pulse remains relatively constant. These features indicate that the extreme events are sporadic, single solitons.

The nonlinear Schrödinger equation (NLSE) models soliton dynamics and has also been used to study hydrodynamic rogue waves generated by nonlinear energy transfer in the open ocean[16–19]. As the NLSE also describes optical pulse propagation in nonlinear media, it is certainly plausible that this equation could predict optical rogue waves. We investigate this numerically using the generalized NLSE (neglecting absorption), which is widely used for broadband optical pulse propagation in nonlinear fibres[9]. The generalized NLSE incorporates dispersion and the Kerr nonlinearity, as well as approximations for self-steepening and the vibrational Raman response of the medium. This equation has been successfully used to model supercontinuum generation in the presence of noise[28,29] and, as we demonstrate here, is capable of qualitatively explaining our experimental results. In anticipation of broadband application, we include several higher orders of dispersion in the nonlinear fibre, which we calculated from the manufacturer's test data (see Methods). Similarly, higher-order dispersion has also been used to extend the validity of the NLSE for broadband calculations in hydrodynamics[30].

As expected, the present model shows that a high-power, smooth input pulse ejects multiple redshifted solitons and blueshifted non-solitonic components, and a tiny amount of input noise varies their spectral content[5,25]. On the other hand, for low power levels, the spectral content of the pulse broadens, but no sharp soliton is shed. In this case, the situation changes markedly when a tiny amount of noise is added. This perturbation is amplified by nonlinear interactions including modulation instability, which dramatically lowers the soliton-fission threshold and permits unpredictable freak events to develop. Interestingly, the hydrodynamic equivalent—the Benjamin–Feir modulation instability—is also thought to initiate hydrodynamic rogue waves[16–19]. This instability spreads spectral content from a narrow bandwidth to a broader range in the initial stages of water wave propagation, just as it does in this optical system.

We include a stochastic perturbation in our simulations by adding to the initial pulse envelope a small amount of bandwidth-restricted random noise with amplitude proportional to the instantaneous field strength. We then solve the NLSE repeatedly for a large number of independent events. For a small fraction of events, the spectrum becomes exceptionally broad with a clear redshifted solitonic shoulder.

Figure 2a shows the time trace and histogram of peak heights for a trial of 1,000 events after red-pass filtering each output pulse at the start of the solitonic shoulder illustrated in Fig. 2b. Clearly, the histogram of heights is sharply peaked but has extended tails, as observed in the experiment, and the distribution contains rogue events more than 50 times as large as the mean. The same rogue events are identified regardless of where the filter is located within the smooth solitonic shoulder and can also be identified from the complementary non-solitonic blue side of the spectrum.

The rogue pulses have exceptionally steep leading and trailing edges compared with the initial pulses and the typical events, as shown in Fig. 2c. The wide bandwidth and abrupt temporal profile of an optical rogue wave is also highlighted in Fig. 3 where the power is displayed as a function of both wavelength and time using a short-time Fourier transform. Because there are no apparent features in the perturbations that lead to the development of the rogue events, their appearance seems unpredictable.

To pinpoint the underlying feature of the noise that produces rogue waves, we have closely analysed the temporal and spectral properties of the initial conditions. Examining the correlations

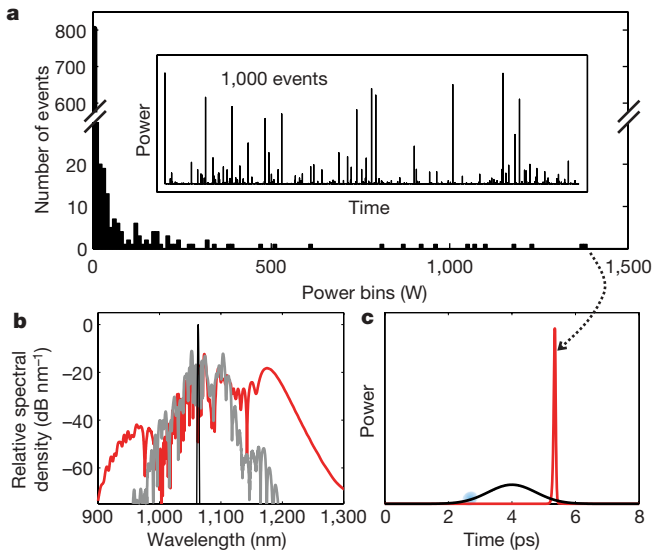


Figure 2 | Simulation of optical rogue waves using the generalized nonlinear Schrödinger equation. **a**, The time trace and histogram of 1,000 events with red-pass filtering from 1,155 nm. The initial (seed) pulses have width 3 ps, peak power 150 W, fractional noise 0.1%, and noise bandwidth 50 THz. The vertical axis of the histogram contains a scale break to make it easier to see the disparity between the most common events at low peak power and the rogue events at high peak power. **b**, The complete relative spectral densities of the initial pulse (black line), a typical event (grey line) and the rare event shown in **c** (red line). **c**, The markedly different temporal profiles of the seed pulse and the rare event indicated in the histogram. The typical events from the histogram are so tiny that they are not visible on this linear power scale. The shaded blue region on the seed pulse delineates the time window that is highly sensitive to perturbation.

between the initial conditions and their respective output waveforms, we find that if the random noise happens to contain energy with a frequency shift of about 8 THz within a 0.5-ps window centred about 1.4 ps before the pulse peak (Fig. 2c), a rogue wave is born. Noise at this particular frequency shift and on a leading portion of the pulse envelope efficiently seeds modulation instability, reshaping the pulse to hasten its breakup. The output wave height correlates in a highly nonlinear way with this specific aspect of the initial conditions. Thus, the normal statistics of the input noise are transformed into an extremely skewed, L-shaped distribution of output wave heights. Further study is needed to explain precisely why the pulse is so highly sensitive to these particular noise parameters. Nevertheless, the specific feature we have identified in the initial conditions offers some predictive power for optical rogue waves, and may offer clues to the oceanic phenomenon.

The rogue waves have a number of other intriguing properties warranting further study. For example, they propagate without noticeable broadening for some time, but have a finite, seemingly unpredictable lifetime before they suddenly collapse owing to cumulative effects of Raman scattering. This scattering seeded by noise dissipates energy or otherwise perturbs the soliton pulse beyond the critical threshold for its survival[9]. The decay parallels the unpredictable lifetimes of oceanic rogue waves. The rogue optical solitons are also able to absorb energy from other wavepackets they pass through, which causes them to grow in amplitude, but appears to reduce their lifetime. A similar effect may help to explain the development of especially large rogues in the ocean.

In conclusion, we have observed extreme soliton-like pulses that are the optical equivalent of oceanic rogue waves. These rare optical events possess the hallmark phenomenological features of oceanic rogue waves—they are extremely large and seemingly unpredictable, follow unusual L-shaped statistics, occur in a nonlinear medium, and are broadband and temporally steep compared with typical events.

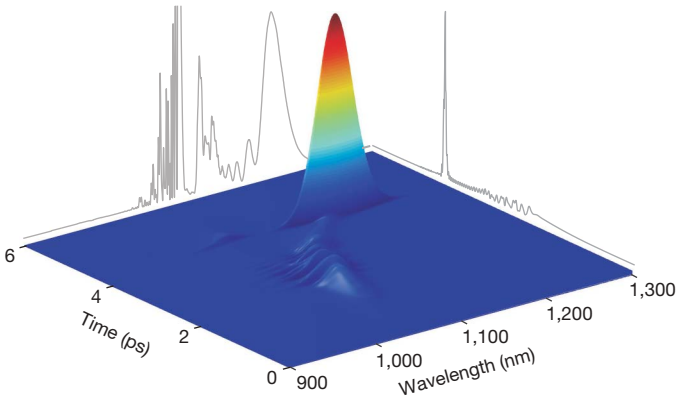


Figure 3 | Time–wavelength profile of an optical rogue wave obtained from a short-time Fourier transform. The optical wave has broad bandwidth and has extremely steep slopes in the time domain compared with the typical events. It appears as a 'wall of light' analogous to the 'wall of water' description of oceanic rogue waves. The rogue wave travels a curved path in time–wavelength space because of the Raman self-frequency shift and group velocity dispersion, separating from non-solitonic fragments and remnants of the seed pulse at shorter wavelengths. The grey traces show the full time structure and spectrum of the rogue wave. The spectrum contains sharp spectral features that are temporally broad and, thus, do not reach large peak power levels and do not appear prominently in the short-time Fourier transform.

On a physical level, the similarities also abound, with modulation instability, solitons, frequency downshifting and higher-order dispersion as striking points of connection. Intriguingly, the rogue waves of both systems can be modelled with the nonlinear Schrödinger equation. Although the parameters that characterize this optical system are of course very different from those describing waves on the open ocean, the rogue waves generated in the two cases bear some remarkable similarities.

METHODS SUMMARY

Our supercontinuum source consists of a master oscillator, a fibre amplifier, and a 15-m length of highly nonlinear microstructured fibre whose zero-dispersion point matches the seed wavelength. The master oscillator is a mode-locked ytterbium-doped fibre laser producing picosecond pulses at about 1,064 nm with a repetition rate of 20 MHz. The output pulses are amplified to a desired level in a large-mode-area ytterbium-doped-fibre amplifier. This amplification process yields chirped pulses of ~5-nm bandwidth and a few picoseconds temporal width.

The wavelength-to-time transformation for real-time detection is accomplished using a highly dispersive optical fibre producing about $-1{,}300$ ps nm^{-1} of GVD over the wavelength range of interest. Because the supercontinuum output is red-pass filtered with a cut-on wavelength of 1,450 nm, adjacent pulses do not overlap in time after being stretched. The GVD-stretched signal is then fed to a fast photodetector and captured by a real-time 20-gigasample-per-second oscilloscope, which records sequences of ~15,000 pulses with high temporal resolution in a single-shot measurement.

The detection of rogue events is insensitive to the filter window, so the specific choice of the red-pass cut-on wavelength is not critical. The soliton shoulder shown in Fig. 2b is smooth and extends to very long wavelengths, so a freak soliton can be detected by examining any section of this extended region. Because of experimental constraints, we limit the measurements to the long-wavelength tail of the soliton shoulder, whereas, in the simulations, it is instructive to capture the entire soliton spectrum. The simulations show that it is acceptable to detect the rogue events experimentally by their red tails because the same rogue events are identified no matter where the filter is located throughout this spectral region.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 22 February; accepted 11 October 2007.

1. Hopkin, M. Sea snapshots will map frequency of freak waves. *Nature* **430**, 492 (2004).

2. Perkins, S. Dashing rogues: freak ocean waves pose threat to ships, deep-sea oil platforms. *Science News* **170,** 328–329 (2006).
3. Broad, W. J. Rogue giants at sea. *The New York Times* (July 11, 2006).
4. Herrmann, J. *et al.* Experimental evidence for supercontinuum generation by fission of higher-order solitons in photonic fibers. *Phys. Rev. Lett.* **88,** 173901 (2002).
5. Dudley, J. M., Genty, G. & Coen, S. Supercontinuum generation in photonic crystal fiber. *Rev. Mod. Phys.* **78,** 1135–1184 (2006).
6. Corwin, K. L. *et al.* Fundamental noise limitations to supercontinuum generation in microstructure fiber. *Phys. Rev. Lett.* **90,** 113904 (2003).
7. Gaeta, A. L. Nonlinear propagation and continuum generation in microstructured optical fibers. *Opt. Lett.* **27,** 924–926 (2002).
8. Alfano, R. R. The ultimate white light. *Sci. Am.* **295,** 87–93 (2006).
9. Agrawal, G. P. *Nonlinear Fiber Optics* 3rd edn (Academic, San Diego, 2001).
10. Kharif, C. & Pelinovsky, E. Physical mechanisms of the rogue wave phenomenon. *Eur. J. Mech. B Fluids* **22,** 603–634 (2003).
11. Dean, R. G. & in. *Water Wave Kinematics* (eds Tørum, A. & Gudmestad, O. T.) 609–612 (Kluwer, Amsterdam, 1990).
12. Muller, P. Garrett, C. & Osborne, A. Rogue waves. *Oceanography* **18,** 66–75 (2005).
13. Walker, D. A. G., Taylor, P. H. & Taylor, R. E. The shape of large surface waves on the open sea and the Draupner New Year wave. *Appl. Ocean. Res.* **26,** 73–83 (2004).
14. Dysthe, K., Socquet-Juglard, H., Trulsen, K., Krogstad, H. E. & Liu, J. "Freak" waves and large-scale simulations of surface gravity waves. *Rogue Waves, Proc. 14th 'Aha Huliko'a Hawaiian Winter Workshop* 91–99 (Univ. Hawaii, Honolulu, 2005).
15. Liu, P. C. & MacHutchon, K. R. Are there different kinds of rogue waves? *Proc. OMAE2006, 25th Int. Conf. Offshore Mechanics and Arctic Engineering*, Paper No. 92619, 1–6 (American Society of Mechanical Engineers, New York, 2006).
16. Henderson, K. L., Peregrine, K. L. & Dold, J. W. Unsteady water wave modulations: fully nonlinear solutions and comparison with the nonlinear Schrödinger equation. *Wave Motion* **29,** 341–361 (1999).
17. Onorato, M., Osborne, A. R., Serio, M. & Bertone, S. Freak waves in random oceanic sea states. *Phys. Rev. Lett.* **86,** 5831–5834 (2001).
18. Onorato, M., Osborne, A. R. & Serio, M. Modulational instability in crossing sea states: A possible mechanism for the formation of freak waves. *Phys. Rev. Lett.* **96,** 014503 (2006).
19. Shukla, P. K., Kourakis, I., Eliasson, B., Marklund, M. & Stenflo, L. Instability and evolution of nonlinearly interacting water waves. *Phys. Rev. Lett.* **97,** 094501 (2006).
20. Janssen, P. A. E. M. Nonlinear four-wave interactions and freak waves. *J. Phys. Oceanogr.* **33,** 863–884 (2003).
21. White, B. S. & Fornberg, B. On the chance of freak waves at sea. *J. Fluid Mech.* **355,** 113–138 (1998).
22. Ranka, J. K., Windeler, R. S. & Stentz, A. J. Visible continuum generation in air–silica microstructure optical fibers with anomalous dispersion at 800 nm. *Opt. Lett.* **25,** 25–27 (2000).
23. Segur, H. *et al.* Stabilizing the Benjamin–Feir instability. *J. Fluid Mech.* **539,** 229–271 (2005).
24. Islam, M. N. *et al.* Femtosecond distributed soliton spectrum in fibers. *J. Opt. Soc. Am. B* **6,** 1149–1158 (1989).
25. Kutz, J. N., Lyngå, C. & Eggleton, B. J. Enhanced supercontinuum generation through dispersion-management. *Opt. Express* **13,** 3989–3998 (2005).
26. Han, Y., Boyraz, O. & Jalali, B. Tera-sample per second real-time waveform digitizer. *Appl. Phys. Lett.* **87,** 241116 (2005).
27. Gu, X. *et al.* Frequency-resolved optical gating and single-shot spectral measurements reveal fine structure in microstructure-fiber continuum. *Opt. Lett.* **27,** 1174–1176 (2002).
28. Nakazawa, M., Kubota, H. & Tamura, K. Random evolution and coherence degradation of a high-order optical soliton train in the presence of noise. *Opt. Lett.* **24,** 318–320 (1999).
29. Boyraz, O., Kim, J., Islam, M. N., Coppinger, F. & Jalali, B. 10 Gb/s multiple wavelength, coherent short pulse source based on spectral carving of supercontinuum generated in fibers. *J. Lightwave Technol.* **18,** 2167–2175 (2000).
30. Trulsen, K. & Dysthe, K. B. A modified nonlinear Schrödinger equation for broader bandwidth gravity waves on deep water. *Wave Motion* **24,** 281–289 (1996).

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to D.R.S. (solli@ucla.edu).

METHODS

Our simulations are based on the nonlinear Schrödinger equation (NLSE), which governs the propagation of optical pulses and has been widely used to model supercontinuum generation in optical fibres. The equation describes the evolution of the slowly varying electric field envelope, $A(z,t)$, in the presence of temporal dispersion and nonlinearity. In its generalized form, the NLSE accounts for dispersion as well as both the electronic (instantaneous) and vibrational (delayed) nonlinearities in silica glass. For many applications, it is sufficient to use approximations for these nonlinearities that are physically intuitive and efficient for numerical computations using the well-known split-step method. Relative to a reference frame co-moving with the optical pulse, this form of the equation can be expressed as

$$\frac{\partial A}{\partial z} - \mathrm{i}\sum_{m=2}\frac{i^m\beta_m}{m!}\frac{\partial^m A}{\partial t^m} = \mathrm{i}\gamma\left[|A|^2A + \frac{\mathrm{i}}{\omega_0}\frac{\partial}{\partial t}\left(|A|^2A\right) - T_\mathrm{R}A\frac{\partial|A|^2}{\partial t}\right]$$

where β_m are values that characterize the fibre dispersion, γ is the nonlinear coefficient of the fibre, ω_0 is the central carrier frequency of the field, and T_R is a parameter that characterizes the delayed nonlinear response of silica fibre[9]. The bracketed terms on the right-hand side of the equation describe the Kerr nonlinearity, self-steepening and the vibrational Raman response of the medium, respectively. For completeness, we include the self-steepening term in our simulations, but we have found that it is not required for rogue wave generation. The Kerr term produces self-phase modulation, and the Raman term causes frequency downshifting of the carrier wave.

This form of the NLSE has been successfully used in the literature to model supercontinuum generation in the presence of noise and is capable of qualitatively explaining our experimental results. In our calculations, we include dispersion up to sixth order, which we calculated from the manufacturer's test data (see Crystal Fibre NL-5.0-1065 for fibre specifications). Operating at the zero dispersion wavelength of the fibre, we use the dispersion parameters $\beta_2 \approx 0$, $\beta_3 \approx 7.67\times10^{-5}\,\mathrm{ps}^3\,\mathrm{m}^{-1}$, $\beta_4 \approx -1.37\times10^{-7}\,\mathrm{ps}^4\,\mathrm{m}^{-1}$, $\beta_5 \approx 3.61\times10^{-10}\,\mathrm{ps}^5\,\mathrm{m}^{-1}$, and $\beta_6 \approx 5.06\times10^{-13}\,\mathrm{ps}^3\,\mathrm{m}^{-1}$. The nonlinear coefficient and the Raman response parameter are given by $\gamma = 11\,\mathrm{W}^{-1}\,\mathrm{km}^{-1}$ and $T_\mathrm{R} = 5\,\mathrm{fs}$. These numbers model the experimental situation, but we find that the NLSE can also produce rogue wave solutions with other values of the parameters.

To generate rogue waves, we perturb the input pulse by adding a very small amount of amplitude noise directly to its temporal envelope. Specifically, at each point in time, a small random number is added to the input field envelope. The noise amplitude at each point is proportional to the instantaneous amplitude of the pulse. The peak power of the unperturbed pulse is chosen to be small enough that the pulse will not break up without the noise perturbation. We then apply a frequency bandpass filter to limit the input noise to a relatively narrow bandwidth around the seed wavelength, sufficient to mimic the optical noise bandwidth of the input field in the experiment. The specific noise amplitude and peak power of the pulse are not critical, but influence the rogue wave generation rate. Noise amplitudes of the order of 0.1% of the pulse amplitude or even significantly less are adequate to observe a rare, but reasonable generation rate. Although we include noise in this particular way, this specific form is not required to create rogue waves—other perturbations produce similar results. This particular form of noise serves as a conceptually simple perturbation that qualitatively accounts for our experimental results. When we solve the NLSE repeatedly given these conditions, rogue waves are produced as statistically rare events from members of an initial population that are nearly indistinguishable.

Vol 450|13 December 2007|doi:10.1038/nature06430

LETTERS

A distinct bosonic mode in an electron-doped high-transition-temperature superconductor

F. C. Niestemski[1], S. Kunwar[1], S. Zhou[1], Shiliang Li[2], H. Ding[1], Ziqiang Wang[1], Pengcheng Dai[2,3] & V. Madhavan[1]

Despite recent advances in understanding high-transition-temperature (high-T_c) superconductors, there is no consensus on the origin of the superconducting 'glue': that is, the mediator that binds electrons into superconducting pairs. The main contenders are lattice vibrations[1,2] (phonons) and spin-excitations[3,4], with the additional possibility of pairing without mediators[5]. In conventional superconductors, phonon-mediated pairing was unequivocally established by data from tunnelling experiments[6]. Proponents of phonons as the high-T_c glue were therefore encouraged by the recent scanning tunnelling microscopy experiments on hole-doped $Bi_2Sr_2CaCu_2O_{8-\delta}$ (BSCCO) that reveal an oxygen lattice vibrational mode whose energy is anticorrelated with the superconducting gap energy scale[7]. Here we report high-resolution scanning tunnelling microscopy measurements of the electron-doped high-T_c superconductor $Pr_{0.88}LaCe_{0.12}CuO_4$ (PLCCO) ($T_c = 24$ K) that reveal a bosonic excitation (mode) at energies of 10.5 ± 2.5 meV. This energy is consistent with both spin-excitations in PLCCO measured by inelastic neutron scattering (resonance mode)[8] and a low-energy acoustic phonon mode[9], but differs substantially from the oxygen vibrational mode identified in BSCCO. Our analysis of the variation of the local mode energy and intensity with the local gap energy scale indicates an electronic origin of the mode consistent with spin-excitations rather than phonons.

Electron- and hole-doped high-T_c superconductors share identical CuO_2 planes where superconductivity originates. Compared with their hole-doped counterparts, the electron-doped copper oxide superconductors represent a largely unexplored territory for scanning tunnelling microscopy (STM) studies where the lack of high-quality samples has posed a tremendous barrier to obtaining high-quality data comparable to that on BSCCO. We have obtained reproducible STM data on nearly optimally doped PLCCO ($T_c = 24$ K) used in recent neutron scattering[8] and angle-resolved photoemission spectroscopy (ARPES) experiments[10]. Figure 1 shows selected STM spectra illustrating the most prominent features in the density of states (DOS): the superconducting gap with coherence peaks, and the step or peak features outside the gap. In addition to these, an obvious feature of the tunnelling spectra is the presence of an almost linear, V-shaped background (Fig. 1b, see also Supplementary Fig. 2) which persists above T_c (Fig. 2d). A similar background was observed in previous STM data on the electron-doped superconductor $Nd_{2-x}Ce_xCuO_4$ (NCCO)[11]. There are several different conjectures for a linear background in the DOS, ranging from momentum-dependent tunnelling matrix element effects in a marginal Fermi liquid[12] to inelastic tunnelling from a continuum of states[13]. Once this background is divided out, however, the muted spectral features (including the formerly suppressed coherence peak heights) come to the forefront as shown in Fig. 1c and d. To make sure that the observed gap is associated with the superconducting gap, we have performed spectroscopy at various temperatures up to 32 K (above T_c) (Fig. 2d) and find that the gap does indeed disappear above T_c. We thus identify the peak-to-peak distance in the local density of states at 5.5 K with twice the local energy gap for superconducting quasiparticles (2Δ).

Although there have been no prior STM studies on PLCCO, ARPES studies[14] point towards a non-monotonic *d*-wave gap with a maximum around 5.5 meV. Point contact tunnelling[15] observes a zero temperature gap ($\Delta(0)$) of 3.5 meV with a ratio $2\Delta(0)/k_BT_c = 3.5 \pm 0.3$ consistent with weak-coupling BCS. Previous STM data obtained on NCCO[11] showed gaps of 3.5 to 5 meV with no obvious coherence peaks. Given the highly inhomogeneous nature of doped layered oxides, spatially resolved STM is a useful key to providing the local energy scales and spatial distribution of the superconducting gap. Statistics of the gap magnitude and its spatial variation were obtained through thousands of spectra (dI/dV mapping) in various regions of the sample (Fig. 2a and c). Although most maps (9 out of 13) reveal average gaps in the range of 6.5–7.0 meV, the average gap (over all measured maps) is 7.2 ± 1.2 meV (Fig. 2c). Approximating $\Delta(0)$ as 7.2 meV allows us to obtain a rough estimate for the ratio $2\Delta/k_BT_c \approx 7.5$, putting the electron-doped superconductors in the strong coupling regime, thereby suggesting a greater overlap between the fundamental physics of the electron- and hole-doped materials than previously shown.

We now turn to important features in the local density of states at energies greater than Δ. A step-like feature in the DOS (which results in a peak in the second derivative of the tunnel current d^2I/dV^2) is normally interpreted as the signature of a bosonic excitation in the system. STM data on bosonic excitations and the strengths of their coupling to the electrons could potentially provide critical information on viable candidates for the pairing mode. Shown in Fig. 3a is a typical dI/dV spectrum obtained on these samples. The derivative of the spectrum (Fig. 3b) reveals peaks at distinct energies marked E_1 and E_2. In the superconducting state, a bosonic excitation appears in STM spectra at an energy offset by the gap, that is, $E = \Omega + \Delta$ where E is the energy of the feature in the spectrum and Ω is the mode energy. Although the line-shape of the feature could be influenced by the details of the process, both inelastic tunnelling effects[16] and electron self-energy effects[6] from a strongly coupled bosonic mode are expected at energies offset by the gap. This allows us conveniently to extract the mode energies for this spectrum ($\Omega_{1,2} = E_{1,2} - 7.0$ meV) resulting in 10.7 meV and 21.7 meV respectively. Because spectral features at multiples of Ω_1 could arise from multi-boson excitations, we find that $\Omega_{1,2}$ are amenable to interpretation as multiples of the same mode Ω_1 at 10.85 ± 0.15 meV.

To determine the statistical significance of the observation of this bosonic mode, we obtained high-resolution dI/dV maps over many

[1]Department of Physics, Boston College, Chestnut Hill, Massachusetts 02467, USA. [2]Department of Physics and Astronomy, The University of Tennessee, Knoxville, Tennessee 37996-1200, USA. [3]Neutron Scattering Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-6393, USA.

regions of the samples and analysed them to extract Δ and E locally. Data from one such map are shown in Fig. 3c. The observation of multiples of Ω_1 allows us to extract Ω_1 in two different ways for each spectrum: $\Omega_1 = E_1 - \Delta$ and $\Omega_1^* = E_2 - E_1$. These are independent observables whose histograms are plotted in Fig. 3d. As can be seen, the two histograms overlap strongly. We thus conclude that the identification of Ω_2 as $2\Omega_1$ bears significant statistical weight, which further supports the identification of these features outside the superconducting gap as originating from bosonic excitations in PLCCO. Using the data from eight dI/dV maps (Fig. 3e), we obtain an average mode energy of $\Omega_{1av} = 10.5 \pm 2.5$ meV. Both the intensity of Ω_1 and the observation of the second harmonic ($2\Omega_1$) of the mode indicate a relatively strong electron-mode coupling. From these spatially resolved spectra, we can also calculate the correlation between the local mode $\Omega(r)$ and the gap $\Delta(r)$. We find that $\Omega(r)$ is anticorrelated with the local gap magnitude $\Delta(r)$ as visible in Fig. 4a. The correlation function obtained between the two is fairly short-ranged, with a normalized on-site ($r = 0$) value close to -0.4, comparable to that found[7] in BSCCO. This anticorrelation is the first indication that this signal arises from an intrinsic excitation rather than an extrinsic inelastic excitation outside the superconducting planes.

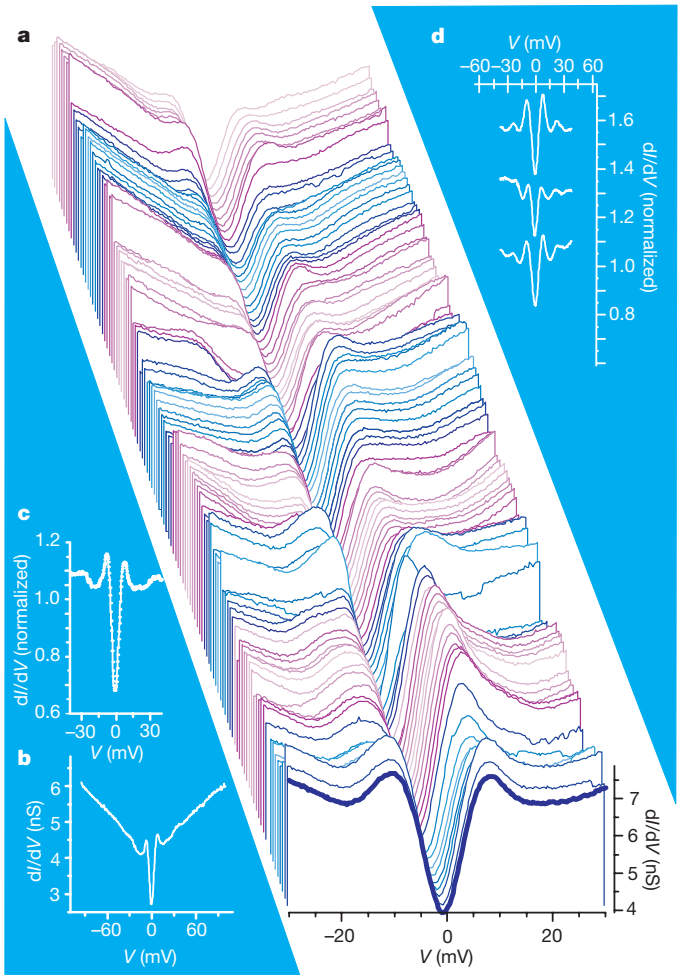


Figure 1 | Prominent low energy spectral features on PLCCO at a temperature of 5.5 K. **a**, A 200-Å section of a 512-Å linecut (a sequence of dI/dV spectra obtained along a spatial line) showing the variations in coherence peak heights and gap magnitude (Δ), defined as half the energy separation between the coherence peaks. The spectra have been offset for clarity. The gap magnitude in this linecut varies from 5 meV to 8 meV. For all spectra, V refers to sample voltage. The spectra were obtained with a junction resistance of 120 MΩ. **b**, A representative ±100-mV range (dI/dV) spectrum (200 MΩ junction resistance) illustrating the dominating V-shaped background. **c**, The linecut reveals spectra that vary from ones with sharp coherence peaks to pseudogap-like spectra without coherence peaks, but most spectra reveal coherence peaks of varying magnitudes once the dominating V-shaped linear background is divided out. This is illustrated by the spectrum shown here, which is the spectrum in **b** after a linear V-shaped division. **d**, More examples of dI/dV spectra demonstrating the clearly resolved coherence peaks and modes resulting from a V-shaped division. These spectra were obtained with 200 MΩ junction resistance.

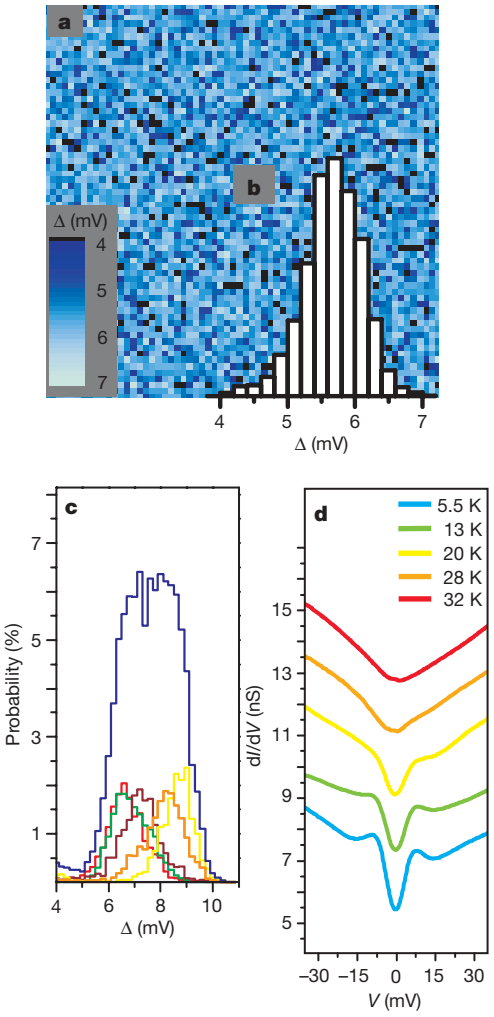


Figure 2 | Gap distribution, statistics and temperature dependence. **a**, A 64 × 64 pixel gap map taken on an area of 256 Å × 256 Å at 5.5 K with a junction resistance of 500 MΩ. The corresponding topographic image reveals no atomic-scale corrugations, as is the common feature of STM images in the superconducting regions of PLCCO (see Supplementary Information), and has therefore been omitted. The patch size in these samples (defined as the area where the gap variation is in the range ±0.5 meV) ranges from 30 Å to 100 Å or more. **b**, The gap distribution of the map in **a**. The average gap in this region is 6.7 ± 1.0 meV. This gap variation is much smaller than that observed in hole-doped BSCCO on a similar-sized region. **c**, Multiple histograms showing gap distributions in different regions of the sample. Our STM's coarse x–y motion capabilities (±0.5 mm maximum) allow us to collect data in regions separated by larger length scales. The mean gap (for the regions represented here) ranges from 6.7 meV to 8.5 meV with a standard deviation ranging from 0.5 meV to 1 meV. Given these statistics, we conclude that gap variations in PLCCO occur on longer length scales than in BSCCO. The sum of these histograms is shown in blue. The average gap over all 13 maps that we obtained is 7.2 ± 1.2 meV. **d**, Temperature evolution of spatially averaged spectra. The spectra have been offset along the vertical axis for clarity. A 256-Å linecut (junction resistance of 120 MΩ) was averaged at each temperature shown, from 5.5 K to 32 K. By 28 K, the gap and coherence peaks have disappeared, but the V-shaped overall background remains.

Having established the mode energy and statistics, and its correlation to the local gap, we now discuss the nature of this excitation. Indeed, the measured mode energy of 10.5 ± 2.5 meV suggests an immediate connection to the 11-meV magnetic resonance mode discovered recently in PLCCO (ref. 8) and NCCO (ref. 17) at $Q = (½, ½, 0)$ by inelastic neutron scattering. The neutron resonance mode, or more precisely its precursor above T_c, has been suggested as a possible pairing glue for the high-T_c copper oxides. Theoretically, bosonic modes originating from spin-excitations can be observed by STM provided there is sufficient coupling between the charge and spin degrees of freedom[18]. Magnetoresistance measurements on underdoped non-superconducting $Pr_{1.3-x}La_{0.7}Ce_xCuO_{4-\delta}$ have provided evidence for strong spin–charge coupling in these materials[19]. It is thus possible that the magnetic resonance mode observed by neutron scattering is related to the observed bosonic mode in the STM signal in PLCCO.

Although magnetic excitations fit the energy scale of our data, another possibility is that the mode originates from in-plane (CuO_2 plane) phonons, like the B_{1g} mode attributed to the STM feature in BSCCO. Compared with BSCCO, however, the energy scale of our mode (10.5 ± 2.5 meV) is much lower. In the hole-doped superconductors a few phonon branches do exist at these low energies[20,21] and the important question is whether there are candidate phonons at these low energies in PLCCO. As it turns out, many of the in-plane phonons in closely related materials, including the B_{1g} mode, have energies higher than 20 meV (refs 22–24) and can therefore be ruled out. Acoustic phonons are viable candidates for this mode provided that the phonon dispersion results in a sharp DOS feature at this energy scale. Such phonons with a DOS peak at or close to 11 meV have indeed been found[9,25] in NCCO. Expanding the search to in-plane phonons at nearby energies reveals an E_u oxygen mode[23] and an oxygen rotation mode[22] at energies greater than 15 meV (15 meV is the lowest energy in the dispersion). We thus conclude that whereas the energy scale of the observed mode clearly rules out the B_{1g} oxygen phonons, at least one in-plane phonon (acoustic) mode does exist at nearby energies.

Apart from these in-plane phonons, the STM mode might arise from inelastic co-tunnelling processes[26] involving an excitation of a local vibrational mode in the intervening layers between the tip and the superconducting plane ('barrier' mode). Such 'out-of-plane'

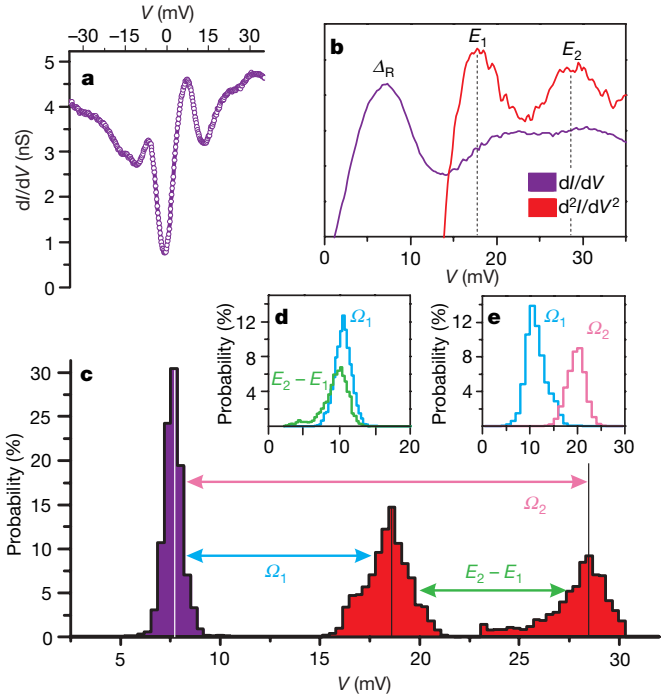


Figure 3 | Statistics of the mode observed as peaks in d^2I/dV^2. **a**, A typical dI/dV spectrum taken at 5.5 K with a junction resistance of 200 MΩ demonstrating the appearance of the modes. **b**, The same spectrum from **a** (purple) as well as its derivative, d^2I/dV^2 (red). The linear V-shaped background has been divided out for clarity and the spectra are now shown only for energies greater than the Fermi energy (E_F). The peak in dI/dV (at 7.0 meV) is the coherence peak, labelled as Δ_R. The peaks in d^2I/dV^2 are labelled as E_1 and E_2 respectively. **c**, A histogram of the occurrences of Δ_R (purple) and the energies E_1 and E_2 (red) for a map of dI/dV on a 64 Å × 64 Å area of the sample. We calculate the average gap (Δ_{av}) in this region to be 7.7 ± 0.5 meV, and the average peaks to be $E_{1av} = 18.5 \pm 1.5$ meV and $E_{2av} \approx 28$ meV (our cut-off at 30 meV for this analysis prevents us from obtaining full statistics for E_2). **d**, Following convention in superconducting systems, the mode energy will be symbolized by Ω ($\Omega_i = E_i - \Delta$). The mode energy is calculated in two ways for each spectrum in the map: $E_1 - \Delta_R$ (blue) with a mean of 10.7 ± 1 meV and $E_2 - E_1$ (green) with a mean of 10 ± 1.7 meV. These are two independent variables and the remarkable overlap between these histograms lends weight to their identification as multiples of the same mode. **e**, Histogram of the mode energies Ω_1 (blue) and Ω_2 (pink) summed for eight maps in different areas of the sample with gaps ranging from 6.5 meV to 8.5 meV. The mode energies were extracted from above and below E_F. From these data, we obtain the average mode energy $\Omega_{1av} = 10.5 \pm 2.5$ meV.

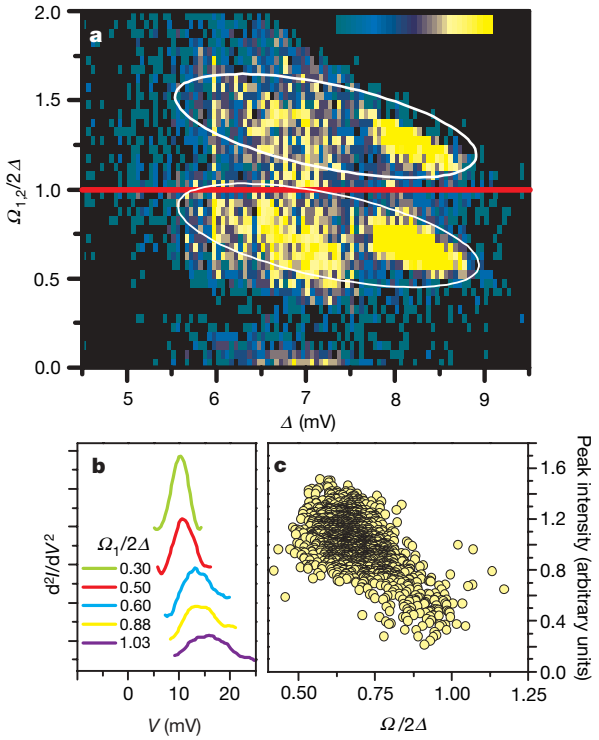


Figure 4 | Variation of local mode energy and intensity with the local gap energy scale. **a**, Log intensity (two-dimensional histogram; blue to yellow shows minimum to maximum) of the occurrences of Ω_1 and Ω_2 plotted as a local ratio $\Omega(r)/2\Delta(r)$ against $\Delta(r)$, clearly revealing the anticorrelation between Ω and Δ. Also note that $\Omega_1(r)/2\Delta(r)$ remains below 1 for a statistically significant part of the data (whereas $\Omega_2(r)/2\Delta(r)$ remains below 2). This demonstrates the sensitivity of the mode to the energy scale 2Δ, which is also borne out by the intensity analysis in **b** and **c**. This plot includes data from three maps obtained in regions of the sample with different average gap values. **b**, Examples of d^2I/dV^2 spectra (from one map) for different ratios of $\Omega_1/2\Delta$ from 0.3 to 1.03. The intensity of the mode (defined as the height of the peak in d^2I/dV^2 spectra) decreases and the mode gets wider in energy as Ω_1 approaches 2Δ. Although both Ω and Δ can vary from spectrum to spectrum, it is the local ratio of $\Omega(r)$ to $2\Delta(r)$ that determines the intensity of the mode. This is consistent with increased damping of the mode associated with the onset of the continuum of excitations at $2\Delta(r)$. **c**, A plot of the mode intensity (reiterating the same behaviour as in **b**) now for all the measured bosonic modes in a single map. Similar intensity drops with the ratio $\Omega_1/2\Delta$ were observed for maps in different regions with average gaps ranging from 6.5 meV to 8.5 meV.

phonons associated with the apical oxygen have been postulated as an alternative explanation for the BSCCO data[16]. Although barrier modes in PLCCO might originate from Pr/La/Ce vibrational excitations in the layers adjacent to the CuO_2 planes, it is not obvious how such modes would lead to our observed correlation between $\Omega(r)$ and $\Delta(r)$. Indeed, based on the idea that this correlation is significant, one recent analysis of the BSCCO STM data[27] postulates two coexisting bosonic modes, only one of which is sensitive to the superconducting gap and can be considered as a signature of the neutron resonance mode in BSCCO.

To bring more insight into the issue of electronic versus lattice-vibrational sources of this mode, we further explore its connection to the superconductivity energy gap. The local nature of STM spectroscopy makes it possible to study the relationship between the spectral properties of the local mode, and the energy scale for the onset of particle–hole excitations ($2\Delta(r)$). Figure 4a is a scatter plot showing the occurrences of the two modes Ω_1 and Ω_2 at a given Δ for three typical regions of different average gap sizes. It is clearly visible in the plot that the ratio $\Omega_1/2\Delta$ lies below 1 for a statistically significant fraction of the observed modes (and that $\Omega_2/2\Delta$ is capped by 2, consistent with the interpretation of Ω_2 as $2\Omega_1$). In Fig. 4b, we present the spectral lines of d^2I/dV^2 near the mode energy for several representative cases with different $\Omega_1/2\Delta$ ratios. Remarkably, the line-shape evolves from a sharp, symmetric feature resembling a resonance peak at low $\Omega_1/2\Delta$ to being broad and asymmetric (like an overdamped mode) as $\Omega_1/2\Delta$ approaches and exceeds one. As shown in Fig. 4c, a clear anticorrelation between the sharpness of the mode and $\Omega_1/2\Delta$ is observed, providing statistical significance for the line-shape analysis in Fig. 4b. These findings unequivocally demonstrate the intimate connection between the bosonic mode and the quasiparticle excitations across the superconducting energy gap. We note that although our analysis of the low-temperature STM data argues against the barrier modes, measurements of the normal-state ($T > T_c$) DOS will provide further, more direct data for the influence of the inelastic barrier co-tunnelling processes on the local tunnelling spectroscopy[28].

The overall picture that finally emerges from our STM studies on PLCCO is the observation of a collective mode in the electronic excitations of the system at 10.5 ± 2.5 meV. Although we cannot rule out low-energy phonons, this mode is fully consistent with the neutron spin resonance mode, and strongly coupled to the superconducting order parameter, making it a compelling candidate boson in the model based on the Eliashberg[29] framework, where exchanging associated electronic (spin or charge) excitations serves as the unconventional pairing mechanism in these materials.

Received 3 June; accepted 18 October 2007.

1. McQueeney, R. J. *et al.* Anomalous dispersion of LO phonons in $La_{1.85}Sr_{0.15}CuO_4$ at low temperatures. *Phys. Rev. Lett.* **82,** 628–631 (1999).
2. Lanzara, A. *et al.* Evidence for ubiquitous strong electron–phonon coupling in high-temperature superconductors. *Nature* **412,** 510–514 (2001).
3. Rossat-Mignod, J. *et al.* Neutron scattering study of the $YBa_2Cu_3O_{6+x}$ system. *Physica C* **185–189,** 86–92 (1991).
4. Norman, M. R. *et al.* Unusual dispersion and line shape of the superconducting state spectra of $Bi_2Sr_2CaCu_2O_{8+\delta}$. *Phys. Rev. Lett.* **79,** 3506–3509 (1997).
5. Anderson, P. W. Is there glue in cuprate superconductors? *Science* **316,** 1705–1707 (2007).
6. McMillan, W. L. & Rowell, J. M. Lead phonon spectrum calculated from superconducting density of states. *Phys. Rev. Lett.* **14,** 108–112 (1965).
7. Lee, J. *et al.* Interplay of electron-lattice interactions and superconductivity in $Bi_2Sr_2CaCu_2O_{8+\delta}$. *Nature* **442,** 546–550 (2006).
8. Wilson, S. D. *et al.* Resonance in the electron-doped high-transition-temperature superconductor $Pr_{0.88}LaCe_{0.12}CuO_{4-\delta}$. *Nature* **442,** 59–62 (2006).
9. d'Astuto, M. *et al.* Anomalous dispersion of longitudinal optical phonons in $Nd_{1.86}Ce_{0.14}CuO_{4+\delta}$ determined by inelastic X-ray scattering. *Phys. Rev. Lett.* **88,** 167002 (2002).
10. Pan, Z.-H. *et al.* Universal quasiparticle decoherence in hole- and electron-doped high-T_c cuprates. Preprint at ⟨http://www.arXiv.org/cond-mat/0610442⟩ (2006).
11. Kashiwaya, S. *et al.* Tunneling spectroscopy of superconducting $Nd_{1.85}Ce_{0.15}CuO_{4-\delta}$. *Phys. Rev. B* **57,** 8680–8686 (1998).
12. Littlewood, P. B. & Varma, C. M. Anisotropic tunneling and resistivity in high-temperature superconductors. *Phys. Rev. B* **45,** 12636 (1992).
13. Kirtley, J. R. & Scalapino, D. J. Inelastic-tunneling model for the linear conductance background in the high-T_c superconductors. *Phys. Rev. Lett.* **65,** 798–800 (1990).
14. Matsui, H. *et al.* Direct observation of a nonmonotonic $d_{x^2-y^2}$-wave superconducting gap in the electron-doped high-T_c superconductor $Pr_{0.89}LaCe_{0.11}CuO_4$. *Phys. Rev. Lett.* **95,** 017003 (2005).
15. Shan, L. *et al.* An universal law of the superconducting gap in the electron-doped cuprate superconductors. *Phys. Rev. B* (in the press). Preprint at ⟨http://www.arXiv.org/cond-mat/0703256⟩ (2007).
16. Pilgram, S., Rice, T. M. & Sigrist, M. Role of inelastic tunneling through the insulating barrier in scanning-tunneling-microscope experiments on cuprate superconductors. *Phys. Rev. Lett.* **97,** 117003 (2006).
17. Zhao, J. *et al.* Neutron-spin resonance in optimally electron-doped superconductor $Nd_{1.85}Ce_{0.15}CuO_4$. *Phys. Rev. Lett.* **99,** 017001 (2007).
18. Zhu, J. X. *et al.* Fourier-transformed local density of states and tunneling into a d-wave superconductor with bosonic modes. *Phys. Rev. B* **73,** 014511 (2006).
19. Lavrov, A. N. *et al.* Spin-flop transition and the anisotropic magnetoresistance of $Pr_{1.3-x}La_{0.7}Ce_xCuO_4$: Unexpectedly strong spin-charge coupling in the electron doped cuprates. *Phys. Rev. Lett.* **92,** 227003 (2004).
20. Pintschovius, L. *et al.* Inelastic neutron scattering study of La_2CuO_4. *Prog. High Temp. Supercond.* **21,** 36–45 (1989).
21. Renker, B. *et al.* Electron–phonon coupling in HTC superconductors evidenced by inelastic neutron scattering. *Physica B* **180,** 450–452 (1992).
22. Pintschovius, L. & Reichardt, W. in *Physical Properties of High Temperature Superconductors* Vol. IV (ed. Ginsberg, D. M.) 295 (World Scientific, London, 1994).
23. Crawford, M. K. *et al.* Infrared active phonons in $(Pr_{2-x}Ce_x)CuO_4$. *Solid State Commun.* **73,** 507–509 (1990).
24. Homes, C. C. *et al.* Optical properties of $Nd_{1.85}Ce_{0.15}CuO_4$. *Phys. Rev. B* **56,** 5525–5534 (1997).
25. Lynn, J. W. *et al.* Phonon density of states and superconductivity in $Nd_{1.85}Ce_{0.15}CuO_4$. *Phys. Rev. Lett.* **66,** 919–922 (1991).
26. Persson, B. N. J. & Baratoff, A. Inelastic electron tunneling from a metal tip: the contribution from resonant processes. *Phys. Rev. Lett.* **59,** 339–342 (1987).
27. Hwang, J., Timusk, T. & Carbotte, J. P. Scanning-tunneling spectra of cuprates. *Nature* **446,** E3–E4 (2007).
28. Scalapino, D. J. Superconductivity: Pairing glue or inelastic tunnelling? *Nature Phys.* **2,** 593–594 (2006).
29. Eliashberg, G. M. Interactions between electrons and lattice vibrations in a superconductor. *Sov. Phys. JETP* **11,** 696–702 (1960).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. V. Balatsky, E. W. Hudson, P. Richard, G. Murthy, J. Engelbrecht and J. C. Davis for discussions and comments. This work was supported by the NSF and the DOE.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to V.M. (madhavan@bc.edu).

Vol 450|13 December 2007|doi:10.1038/nature06382

LETTERS

Infinite-layer iron oxide with a square-planar coordination

Y. Tsujimoto[1], C. Tassel[1,2], N. Hayashi[3], T. Watanabe[1], H. Kageyama[1], K. Yoshimura[1], M. Takano[4,5], M. Ceretti[2], C. Ritter[6] & W. Paulus[2]

Conventional high-temperature reactions limit the control of coordination polyhedra in transition-metal oxides to those obtainable within the bounds of known coordination geometries for a given transition metal[1]. For example, iron atoms are almost exclusively coordinated by three-dimensional polyhedra such as tetrahedra and octahedra. However, recent works have shown that binary metal hydrides act as reducing agents at low temperatures, allowing access to unprecedented structures[2–4]. Here we show the reaction of a perovskite $SrFeO_3$ with CaH_2 to yield $SrFeO_2$, a new compound bearing a square-planar oxygen coordination around Fe^{2+}. $SrFeO_2$ is isostructural with 'infinite layer' cupric oxides[5–8], and exhibits a magnetic order far above room temperature in spite of the two-dimensional structure, indicating strong in-layer magnetic interactions due to strong Fe *d* to O *p* hybridization. Surprisingly, $SrFeO_2$ remains free from the structural instability that might well be expected at low temperatures owing to twofold orbital degeneracy in the Fe^{2+} ground state with D_{4h} point symmetry. The reduction and the oxidation between $SrFeO_2$ and $SrFeO_3$ proceed via the brownmillerite-type intermediate $SrFeO_{2.5}$, and start at the relatively low temperature of ~400 K, making the material appealing for a variety of applications, including oxygen ion conduction, oxygen gas absorption and catalysis.

The coordination number in ionically bonded structures is governed by the relative size of oppositely charged ions, which to a first approximation may be regarded as charged rigid spheres (Pauling's first rule)[1]. Each ion 'tries' to surround itself symmetrically with the largest possible number of oppositely charged neighbours. Transition metal ions normally have, owing to their relatively small ionic radii, a preference for tetrahedral and octahedral coordination, as found typically in perovskites and spinels.

For an octahedral coordination there is a pronounced effect on the stereochemistry, or a substantial deviation from ideal geometry, for the Jahn–Teller ions with $(e_g)^1$ or $(e_g)^3$ configuration, such as Cr^{2+}, $Mn^{3+}(d^4)$ and Cu^{2+}, $Ni^{+}(d^9)$, in which the lengthening of a pair of bonds perpendicular to the equatorial plane gives a tetragonal distortion ($c/a > 1$) and results ultimately in a square-planar coordination[1]. Among many copper oxides with square planar geometry, the series $ACuO_2$ (where A = Sr, Ca)[5,6] are known as the "infinite-layer compounds", or the mother structure of high-transition-temperature (T_c) superconductors[7,8], which consist of a sequence of infinitely repeating stacking of CuO_2 square lattices. The isostructural phase of $LaNiO_2$ is formed naturally because monovalent nickel is isoelectronic with divalent copper[9]. Unusual coordination beyond the constraint of the ionic model may occur in association with covalent bonding, which is highly directional, as realized, for example, in many silicates and sulphides[10]. A successful approach for exploring unusual coordinations in molecular compounds is the use of organic ligands with specific steric constraints[11,12]. This cannot be applied, however, to nonmolecular inorganic solids.

Iron, one of the richest elements in the Earth, forms an uncountable number of oxides, some of which have been widely used in industry as low-cost ferrite magnets and pigments. In almost all of them, the iron ions are tetrahedrally or octahedrally coordinated. To the best of our knowledge, the only example of square-planar coordination is represented by the mineral gillespite $BaFeSi_4O_{10}$ (ref. 13). However, the iron atoms in this oxide are dispersed separately within the building blocks made of four-membered rings of SiO_4^{4-} tetrahedra, rendering the electromagnetic properties of minor interest. Ordered perovskites $CaFe_3Ti_4O_{12}$ and $CaFeTi_2O_6$ have also been discussed as synthetic examples of the square planar geometry[14], but they actually have four short (~2.0 Å) and four long (~2.8 Å) Fe–O bonds, resulting in a three-dimensional environment.

Here we show a low-temperature route for the topotactic synthesis of a new infinite-layer iron oxide, $SrFeO_2$ (Fig. 1b), using an easy-to-prepare, slightly oxygen-deficient perovskite $SrFeO_{3-y}$ ($y \approx 0.125$) as a precursor (Fig. 1a). The formation of the strongly anisotropic framework of $SrFeO_2$ is remarkable in that (1) the Jahn–Teller effect should not be of essential importance for divalent iron ions, (2) such an unusual geometry is obtained for a simple ionic compound, and (3) unlike $BaFeSi_4O_{10}$, the FeO_2 layers form the primary building blocks, making it a quasi-two-dimensional magnet. Furthermore, the oxygen content in $AFeO_{3-y}$ (A = Sr, Ca) has been taken for granted to be in the range of only $0 \leq y \leq 0.5$, and thus the brownmillerite phase $SrFeO_{2.5}$ ($y = 0.5$) (see Supplementary Fig. 1), consisting of alternative layers of FeO_6 octahedra and FeO_4 tehrahedra, was historically assumed to represent the lower limit of oxygen stoichiometry. This is in spite of extensive experimental efforts to control oxygen content that include synthesis at high temperature with oxygen partial pressures ranging from 10^{-9} to 100 MPa (refs 15, 16), electrochemical reactions in aqueous solution at room temperature[17], and the fabrication of epitaxial films under different atmospheres[18].

Our approach follows the recent success in the use of hydrides of electropositive metals as powerful reducing agents[2–4,19–22] such as NaH and CaH_2. For example, using NaH yields $LaSrCoO_{3.38}$ from $LaSrCoO_4$ (ref. 2) and $YSr_2Mn_2O_{5.5}$ from $YSr_2Mn_2O_7$ (ref. 19), and using CaH_2 yields $LaSrCoO_3H_{0.7}$ from $LaSrCoO_4$ (ref. 3), $Yb_2Ti_2O_{6.43}$ from $Yb_2Ti_2O_7$ (ref. 4), and $La_3Ni_2O_6$ from $La_3Ni_2O_7$ (ref. 20). These metal hydrides, routinely used as drying agents in organic synthesis, are now regarded as promising reducing agents for nonmolecular compounds to yield unusual frameworks and coordinations in nonmolecular solids, because they are active at

[1]Department of Chemistry, Graduate School of Science, Kyoto University, Sakyo, Kyoto 606-8502, Japan. [2]University of Rennes 1, Sciences Chimiques de Rennes UMR CNRS 6226, Campus de Beaulieu Bâtiment 10B, Rennes cedex 35042, France. [3]Graduate School of Human and Environmental Studies, Kyoto University, Sakyo, Kyoto 606-8501, Japan. [4]Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan. [5]Research Institute for Production Development, 15 Morimoto, Shimogamo, Sakyo, Kyoto 606-0805, Japan. [6]Institute Laue Langevin, BP 156, 38042, Grenoble, France.

considerably lower temperatures than when conventional techniques are used. They also allow solution chemistry to be avoided. However, their potential for the synthesis of new nonmolecular compounds still seems far from realization.

The powder X-ray diffraction (XRD) patterns of the precursor phase $SrFeO_{2.875}$ (that is, $Sr_8Fe_8O_{23}$) were assigned as the nearly cubic perovskite phase with $a_p \approx 3.86$ Å, though it in fact had a tetragonal $2a_p\sqrt{2} \times 2a_p\sqrt{2} \times 2a_p$ supercell ($I4/mmm$, $a = 10.929$ Å and $c = 7.698$ Å), consistent with the literature[15,16]. On the other hand, the synchrotron XRD patterns of the final product (Supplementary Fig. 2) were readily indexed assuming the tetragonal unit cell to have $a = 3.99107(3)$ Å and $c = 3.47481(5)$ Å, which is completely different from those of any reported $SrFeO_{3-y}$ ($0 \leq y \leq 0.5$) phases. No extra diffraction lines were detected. Compared with the lattice parameters ($a_p \approx 3.86$ Å) of the precursor, the a axis of the final product is slightly increased, but the c axis is drastically decreased. This implies an anisotropic extraction of oxygen atoms located originally on the c axis.

Furthermore, the similarity of the lattice parameters to those of $SrCuO_2$ ($a = 3.926$ Å and $c = 3.432$ Å; ref. 6) and also the fact that no specific extinction rules for reflections could be determined from the diffraction pattern, strongly suggest that the two phases are isostructural in the space group of $P4/mmm$. The Rietveld refinement from the synchrotron data immediately converged to $R_{wp} = 6.04\%$ and $\chi^2 = 4.41$ along with reasonable individual, isotropic displacement factors for all the atoms, suggesting successful structural analysis. The refinement of the site occupancy factor for oxygen atoms gave 0.994(8), indicating that the occupation of the oxygen ($2f$) position is unity within the standard deviation. Inclusion of oxygen atoms into the apical ($1b$) site did not improve the refinement. The bond valence sum calculations gave valences of +1.92 for Sr and +1.97 for Fe, which are in excellent agreement with the expected valences of +2 for both. The elemental analysis by energy dispersive spectroscopy (EDS) before and after the reduction gave the same molar ratio of Sr:Fe = 1:1.

The neutron powder diffraction (NPD) patterns of $SrFeO_2$ at 293 K (Fig. 2a) confirmed the above structure with an excellent convergence ($R_{wp} = 4.70\%$ and $\chi^2 = 3.18$). They also excluded a possible incorporation of hydrogen into the structure, and revealed the presence of the (π, π, π) antiferromagnetic order, where the magnetic moments are perpendicular to the c axis (Fig. 2b). This is the same spin structure as that of the undoped, antiferromagnetically ordered mother phase of high-T_c copper oxide superconductors. The magnetic moment has been found to be $3.1\mu_B$ per Fe atom at 293 K and $3.6\mu_B$ at 10 K (Supplementary Fig. 3). The magnitude itself and its minor variation over the wide temperature range strongly suggest that the ferrous ions are in the high-spin state of $(d_{xz}, d_{yz})^3(d_{xy})^1(d_{z^2})^1(d_{x^2-y^2})^1$ with $S = 2$ (where S is the magnitude of spin) and that the antiferromagnetic transition temperature is considerably higher than room temperature. The magnetic order at ambient temperature was also confirmed by ^{57}Fe Mössbauer spectroscopy at 285 K (Fig. 3b) showing six well-defined peaks with a fairly large hyperfine field of 40.1 T.

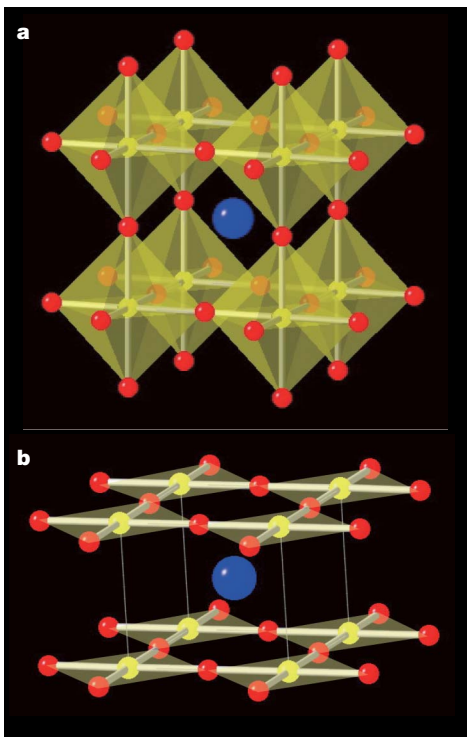


Figure 1 | Structural transformation via a topotactic route. **a**, The cubic perovskite $SrFeO_3$. **b**, The infinite-layer compound $SrFeO_2$. Iron, strontium and oxygen atoms are represented as yellow, blue and red spheres, respectively. For the sake of simplicity, an idealized and stoichiometric cubic phase, obtainable under high oxygen pressure, is demonstrated in **a**, instead of the distorted, slightly oxygen-deficient phase $SrFeO_{3-y}$ ($y \approx 0.125$) used in this study. The iron coordinations by oxygen atoms are illustrated by octahedra in **a** and squares in **b**.

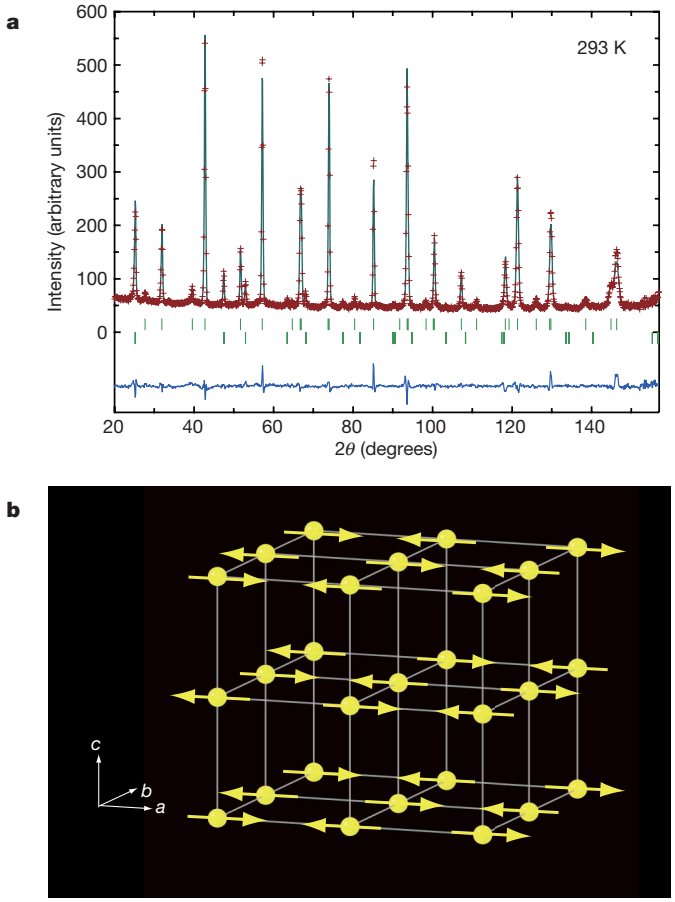


Figure 2 | Structural characterization of $SrFeO_2$ by Rietveld refinement of high-resolution neutron diffraction at room temperature. **a**, The solid lines and overlying crosses indicate the calculated and observed intensities. The green lines are the positions of the calculated nuclear (top) and magnetic (bottom) Bragg reflections. The difference between the observed and calculated profiles is plotted at the bottom in blue. $SrFeO_2$ adopts the $P4/mmm$ space group, $a = 3.991(1)$ Å, $c = 3.474(1)$ Å, Sr on $1d$ (0.5, 0.5, 0.5), Fe on $1a$ (0, 0, 0) and O on $2f$ (0.5, 0, 0), with 100% occupancy, $B_{iso}(Sr) = 0.47(5)$ Å^2, $B_{iso}(Fe) = 0.47(4)$ Å^2, $B_{iso}(O) = 0.79(5)$, $R_p = 3.82\%$, $R_{wp} = 4.70\%$, $\chi^2 = 3.18$, $R_{Bragg} = 4.19\%$, $R_{mag} = 9.19\%$ (see Methods for definitions). **b**, The magnetic structure with a $2a_p \times 2a_p \times 2c_p$ magnetic unit cell of sides of length a, b and c, where iron sites are drawn. Arrows denote the direction of the magnetic moment.

The temperature variations of the magnetic NPD peak intensity and the hyperfine field nicely coincide with each other (see Fig. 3a and c, and Supplementary Table 1), giving the Néel temperature $T_N = 473$ K. The quadrupole interaction appearing in the magnetically split Mössbauer spectrum, which is equal to $S_1 - S_2$ in Fig. 3b, is almost temperature-independent (for example, 1.16 mm s^{-1} at 285 K), and is very close to the quadrupole splitting in the paramagnetic state, ΔE (Fig. 3b), of 1.06 mm s^{-1}. This proximity and the fact that the electric field gradient is uniaxial along the *c* axis by symmetry indicated that the magnetic hyperfine field lies in the *a*–*b* plane[23], consistent with the NPD results.

In spite of the apparent two-dimensionality in magnetism, the T_N is considerably higher than that (~200 K) of FeO having the rock-salt-type, three-dimensionally extended linear Fe–O–Fe bonding. This shows that $SrFeO_2$ has fairly large in-plane exchange constants owing to strong in-plane Fe $d_{x^2-y^2}$ to O $p_{x,y}$ hybridization. Another aspect of the strong hybridization is the isomer shift (~0.5 mm s^{-1} at room temperature), which is located in the extreme covalent limit for a high-spin divalent iron[23]. We also point out that the electronic configuration should be a twofold orbital degenerate ground state of $(d_{xz}, d_{yz})^3$ (ref. 1), because all iron atoms in $SrFeO_2$ are in a high-spin state with D_{4h} point symmetry. Remarkably, $SrFeO_2$ is free from instabilities such as orbital ordering or Jahn–Teller distortion even at the temperatures of 10 and 4.2 K, as derived from NPD and Mössbauer spectroscopy, respectively. We believe that the orbital instability is overcome by the extremely strong covalency that favours directional and symmetrical Fe–O bonding. Therefore, by applying the present synthetic approach to other iron oxides, such as conventional ferrites, we should be able to obtain novel magnetic materials containing Fe^{2+} ions in square-planar geometry and thus having greater magnetic anisotropy and higher transition temperatures.

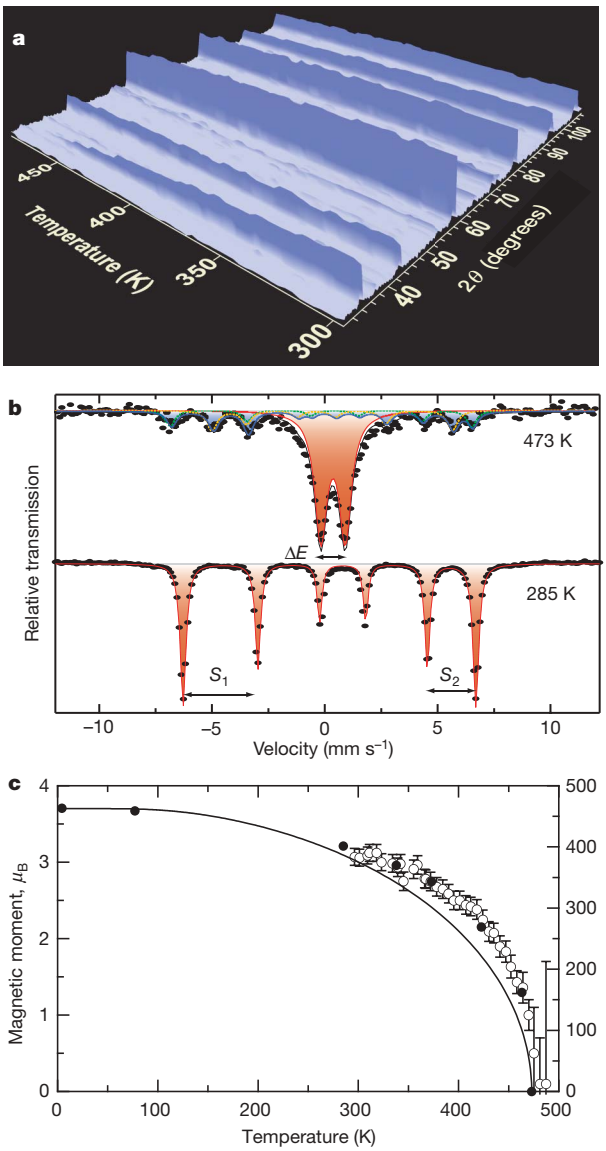


Figure 3 | Temperature evolution of the magnetic order in $SrFeO_2$. **a**, The *in situ* NPD profiles upon warming. The peak at $2\theta = 36^\circ$ corresponds to the (1/2, 1/2, 1/2) magnetic reflection. **b**, Mössbauer spectra at 285 and 473 K. The red lines indicate the spectra from $SrFeO_2$, while the blue line at 473 K represents the spectrum from a small amount of $SrFeO_{2.5}$. The black line is the total fit. **c**, Temperature dependence of the magnetic moment and the hyperfine field H_{hf} determined by NPD (open circles) and Mössbauer spectroscopy (solid circles), respectively. The solid curve is the theoretical Brillouin function for $S = 2$. Error bars are s.d., determined by the Rietveld refinement.

$SrFeO_{3-y}$ and related iron perovskite oxides have been intensely studied, because they exhibit fast oxygen transport combined with high electron conductivity even at low temperatures[17]. They are thus potential candidates for applications as electrodes for solid oxide fuel cells and batteries[24], membranes for oxygen separation[25] and electrocatalysis[26] and gas sensors[27]. In addition, their unusual properties include aspects of vacancy orderings (at $y = 0$, 0.125, 0.25 and 0.5)[15–17], charge disproportionation[28], giant magnetoresistance[29], helical antiferromagnetic spin structure[30], and high-spin to low-spin transition[28]. Here we have shown that the reduction of $SrFeO_{2.875}$ does not necessarily stop at the brownmillerite structure $SrFeO_{2.5}$, but goes beyond this stoichiometry to form a terminal phase with an infinite number of FeO_2 layers. This means that CaH_2 can provide a balanced reducing potential high enough to produce $SrFeO_2$ but not so high that it would result in over-reduction to Fe metal. Starting from $SrFeO_{3-y}$, a one-day reaction in the ranges 473 K $< T <$ 523 K, 523 K $< T <$ 673 K and 673 K $< T$ produces $SrFeO_{2.5}$, $SrFeO_2$ and SrO–Fe mixture, respectively, whereas a one-week reaction at 473 K yields only $SrFeO_2$, demonstrating ever more control of reducing power by varying synthetic temperature and time.

We note that the reduction from $SrFeO_{2.5}$ to $SrFeO_2$ is not a naive topotactic reduction because it involves the filling of the originally vacant sites within the tetrahedral layers of $SrFeO_{2.5}$. Upon heating in an oxygen atmosphere of 0.1 MPa, an opposite reaction back up to $SrFeO_{2.875}$ via $SrFeO_{2.5}$ takes place (Supplementary Fig. 4). Surprisingly, both the reduction and the oxygen uptake proceed at temperatures as low as ~400 K. This implies not only that oxygen is highly mobile in solids at low temperatures, but also that a given dense framework can rearrange towards new oxygen-ordered structures, which may be useful for solid oxide fuel cells, oxygen membranes and sensor materials oriented towards the reduction of working temperatures. It may also be of interest to investigate hole- or electron-doping into the infinite-layer FeO_2 sheets—for example, by introducing selectively apical oxygen atoms or by replacing Sr sites with a monovalent metal like Na.

METHODS SUMMARY

The reduction of $SrFeO_{2.875}$ was performed using CaH_2 as a reducing agent, as described for $LaSrCoO_4$ (ref. 3). $SrFeO_{2.875}$ and a two-molar excess of CaH_2 were finely ground in an Ar-filled glove box, sealed in an evacuated Pyrex tube, and reacted at 553 K for two days. The residual CaH_2 and the CaO byproduct were removed from the final reaction phase by washing them out with a NH_4Cl/methanol solution. Chemical analyses were based on EDS and thermogravimetry. The synchrotron powder XRD experiment was performed on the large Debye–Scherrer camera installed at the SPring-8 beam line BL02B2 by using an imaging plate as a detector. The wavelength of the X-ray is 0.77747 Å. *Ex situ* NPD studies were carried out on the D1A diffractometer, installed at the Institute Laue Langevin. The wavelength of $\lambda = 1.91$ Å was used. The *in situ* NPD profiles upon warming in dynamic vacuum was carried out on the D1B diffractometer installed at the Institute Laue Langevin, where $\lambda = 2.52$ Å was used. ^{57}Fe Mössbauer spectra were taken under dynamical vacuum using a ^{57}Co/Rh source and a control absorber of α-Fe.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 3 July; accepted 5 October 2007.

1. Wells, A. F. *Structural Inorganic Chemistry* 3rd edn (Oxford Univ. Press, Oxford, UK, 1962).
2. Hayward, M. A. & Rosseinsky, M. J. Anion vacancy distribution and magnetism in the new reduced layered Co(II)/Co(I) phase $LaSrCoO_{3.5-x}$. *Chem. Mater.* **12,** 2182–2195 (2000).
3. Hayward, M. A. *et al.* The hydride anion in an extended transition metal oxide array: $LaSrCoO_3H_{0.7}$. *Science* **295,** 1882–1884 (2002).
4. Blundred, G. D., Bridges, A. B. & Rosseinsky, M. J. New oxidation states and defect chemistry in the pyrochlore structure. *Angew. Chem. Intl Edn* **43,** 3562–3565 (2004).
5. Siegrist, T., Zahurak, S. M., Murphy, D. W. & Roth, R. S. The parent structure of the layered high-temperature superconductors. *Nature* **334,** 231–232 (1988).
6. Takano, M., Takeda, Y., Okada, H., Miyamoto, M. & Kusaka, T. $ACuO_2$ (A: alkaline earth) crystallizing in a layered structure. *Physica C* **159,** 375–378 (1989).
7. Smith, M. G., Manthiram, A., Zhou, J. & Goodenough, J. B. Electron-doped superconductivity at 40 K in the infinite-layer compound $Sr_{1-y}Nd_yCuO_2$. *Nature* **351,** 549–551 (1991).
8. Azuma, M., Hiroi, Z., Takano, M., Bando, Y. & Takeda, Y. Superconductivity at 110 K in the infinite-layer compound $(Sr_{1-x}Ca_x)_{1-y}CuO_2$. *Nature* **356,** 775–776 (1992).
9. Crespin, M., Levitz, P. & Gatineau, L. Reduced forms of $LaNiO_3$ perovskite. 1. Evidence for new phases: $La_2Ni_2O_5$ and $LaNiO_2$. *J. Chem. Soc. Faraday Trans.* **2,** 1181–1194 (1983).
10. Hyde, B. G. & Andersson, S. *Inorganic Crystal Structure* Ch. 15 (John Wiley & Sons, New York, 1989).
11. Berry, J. F. *et al.* An octahedral coordination complex of iron(VI). *Science* **312,** 1937–1941 (2006).
12. Bouwkamp, M. W., Bowman, A. C., Lobkovsky, E. & Chirik, P. J. Iron-catalyzed [$2\pi + 2\pi$] cycloaddition of α,ω-dienes: the importance of redox-active supporting ligands. *J. Am. Chem. Soc.* **128,** 13340–13341 (2006).
13. Hazen, R. M. & Burnham, C. W. The crystal structures of gillespite I and II: a structural determination at high pressure. *Am. Mineral.* **59,** 1166–1176 (1974).
14. Leinenweber, K., Linton, J., Navrotsky, A., Fei, Y. & Parise, J. B. High-pressure perovskites on the join $CaTiO_3$-$FeTiO_3$. *Phys. Chem. Mineral.* **22,** 251–258 (1995).
15. Takeda, Y. *et al.* Phase relation in the oxygen nonstoichiometric system $SrFeO_x$ ($2.5 \leq x \leq 3$). *J. Solid-State Chem.* **63,** 237–249 (1986).
16. Hodges, J. P. *et al.* Evolution of oxygen-vacancy ordered crystal structures in the perovskite series $Sr_nFe_nO_{3n-1}$ ($n = 2$, 4, 8, and ∞), and the relationship to electronic and magnetic properties. *J. Solid-State Chem.* **151,** 190–209 (2000).
17. Grenier, J.-C. *et al.* Electrochemical oxygen intercalation into oxide networks. *J. Solid-State Chem.* **96,** 20–30 (1992).
18. Hayashi, N., Terashima, T. & Takano, M. Oxygen-holes creating different electronic phases in Fe^{4+}-oxides: successful growth of single crystalline films of $SrFeO_3$ and related perovskites at low oxygen pressure. *J. Mater. Chem.* **11,** 2235–2237 (2001).
19. Hayward, M. A. Structural and magnetic properties of topotactically reduced $YSr_2Mn_2O_{7-x}$ ($0 < x < 1.5$). *Chem. Mater.* **18,** 321–327 (2006).
20. Poltavets, V. V. *et al.* $La_3Ni_2O_6$: a new double T′-type nickelate with infinite $Ni^{1+/2+}O_2$ layers. *J. Am. Chem. Soc.* **128,** 9050–9051 (2006).
21. Hayward, M. A. Phase separation during the topotactic reduction of the pyrocholore $Y_2Ti_2O_7$. *Chem. Mater.* **17,** 670–675 (2005).
22. Hayward, M. A., Green, M. A., Rosseinsky, M. J. & Sloan, J. Sodium hydride as a powerful reducing agent for topotactic oxide deintercalation: synthesis and characterization of the nickel(I) $LaNiO_2$. *J. Am. Chem. Soc.* **121,** 8843–8854 (1999).
23. Greenwood, N. N. & Gibb, T. C. *Mössbauer Spectroscopy* Ch. 3 & 5 (Chapman and Hall, London, 1971).
24. Shao, Z. & Haile, S. M. A high-performance cathode for the next generation of solid-state fuel cells. *Nature* **431,** 170–173 (2004).
25. Sammells, A. F., Schwartz, M., Mackay, R. A., Barton, T. F. & Peterson, D. R. Catalytic membrane reactors for spontaneous synthesis gas production. *Catal. Today* **56,** 325–328 (2000).
26. Badwal, S. P. S. & Ciacchi, F. T. Ceramic membrane technologies for oxygen separation. *Adv. Mater.* **13,** 993–996 (2001).
27. Wang, Y., Chen, J. & Wu, X. Preparation and gas-sensing properties of perovskite-type $SrFeO_3$ oxide. *Mater. Lett.* **49,** 361–364 (2001).
28. Takano, M. *et al.* Pressure-induced high-spin to low-spin transition in $CaFeO_3$. *Phys. Rev. Lett.* **67,** 3267–3270 (1991).
29. Battle, P. D. *et al.* Magnetoresistance in high oxidation state iron oxides. *Chem. Commun.* **767,** 987–988 (1998).
30. Mostovoy, M. Helicoidal ordering in iron perovskite. *Phys. Rev. Lett.* **94,** 137205 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank Y. Kiuchi, H. Ueda, M. Isobe and Y. Ueda for their help in EDS and thermogravimetric measurements and K. Kato for his help in the synchrotron X-ray experiments at SPring-8. This work was supported by Young Scientists A (H.K.), the Grant-in-Aid for Scientific Research on Priority Areas (H.K. and K.Y.) and Scientific Research S (M.T.) from MEXT. See the Supplementary Notes for more details.

Author Contributions H.K. designed the study in collaboration with W.P, with M.T.'s help; C.T. performed the initial synthesis and proposed the structural model; Y.T. and T.W. optimized the synthetic conditions, performed chemical characterizations, X-ray diffraction measurements and corresponding structural refinement; N.H. conducted the Mössbauer experiment, with M.T.'s help; M.C., C.R. and W.P. performed the neutron diffraction measurements and M.C. and W.P. performed the corresponding structural refinement; All the authors discussed the results; H.K. wrote the manuscript, with comments mainly from M.T. and W.P.

Author Information Atomic coordinates and structure factors for the crystal structure of $SrFeO_2$ have been deposited with the ICSD database under accession codes 418603, 418605 and 418606. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to H.K. (kage@kuchem.kyoto-u.ac.jp).

METHODS

A precursor $SrFeO_{2.875}$ ($y \approx 0.125$) was prepared by a conventional high-temperature ceramic method from predried $SrCO_3$ (99.99%) and Fe_2O_3 (99.99%). Stoichiometric amounts of $SrCO_3$ and Fe_2O_3 were ground together, heated at 1,273 K in air, ground again, and heated for an additional 24 h at 1,473 K. The reduction of $SrFeO_{2.875}$ was performed using CaH_2 as a reducing agent. $SrFeO_{2.875}$ (0.45 g) and a two-molar excess of CaH_2 (0.2 g) were finely ground in an Ar-filled glove box, sealed in an evacuated Pyrex tube (volume 15 cm^3) with a residual pressure less than 1.3×10^{-8} MPa, and reacted at 553 K for two days. The residual CaH_2 and the CaO byproduct were removed from the final reaction phase by washing them out with 0.1 M NH_4Cl in dried methanol.

The EDS experiments were carried out for the precursor and the final product using a JEOL (JSM-5600) scanning electron microscope equipped with an EDAX (Oxford Link ISIS) microanalytical system. Thermogravimetric measurements were performed using a Mac Science thermal analyser (TG-DTA2000). Measurements to analyse re-oxidation behaviour of $SrFeO_2$ were performed on a sample of around 20 mg that was rapidly loaded into an aluminium crucible and then heated at 10 K min^{-1} under flowing O_2 (0.1 MPa). Before the experiment, the sample was dried at 373 K for about 1 hour. The identity and oxygen stoichiometry of the re-oxidized products were determined by XRD.

The synchrotron powder diffraction experiment was performed on the large Debye–Scherrer camera installed at SPring-8 BL02B2 by using an imaging plate as a detector. Incident beams from a bending magnet were monochromatized to 0.77747 Å. The sample was contained in a glass capillary tube with an inner diameter of 0.1 mm and was rotated during measurements. The synchrotron X-ray diffraction data were collected at room temperature in a 2θ range from 1° to 75° with a step interval of 0.01°.

The *ex situ* neutron powder diffraction studies were carried out on the D1A diffractometer, installed at the Institute Laue Langevin (Grenoble, France). A 200 mg sample sealed in a He-filled vanadium can was used and a wavelength of $\lambda = 1.91$ Å was used. The *in situ* neutron powder diffraction experiments of $SrFeO_2$ upon warming in dynamic vacuum ($p_{O_2} < 1.3 \times 10^{-10}$ MPa) were carried out using the Institute Laue Langevin vacuum furnace on the D1B diffractometer installed at Institute Laue Langevin, where $\lambda = 2.52$ Å was used. The temperature was varied from 270 to 500 K with a 0.5 K min^{-1} ramp rate. The temperature stability and average temperature were determined using a thermocouple placed at the sample position.

Mössbauer spectra of $SrFeO_2$ were taken under a dynamical vacuum, and the data were collected in transmission geometry by using a ^{57}Co/Rh γ-ray source at low temperature in combination with a constant-acceleration spectrometer. The source velocity was calibrated by using pure α-Fe as a control material. The low temperature measurements were carried out using a cryostat, while the high-temperature spectra were taken with a small amount of CaH_2 that was mixed with the $SrFeO_2$ powder to minimize re-oxidation of $SrFeO_2$ into the brownmillerite phase $SrFeO_{2.5}$ upon being heated. The obtained spectra were fitted by a lorentzian function. The subspectrum of $SrFeO_{2.5}$ observed at high temperatures was compared with that of ref. 31.

The X-ray and neutron diffraction patterns were analysed by the Rietveld method using the RIETAN 2000 (ref. 32) and FULLPROF software[33], respectively. The agreement indices used were the profile, $R_p = \sum|y_{io} - y_{ic}|/\sum y_{io}$, weighted profile, $R_{wp} = [\sum w_i(y_{io} - y_{ic})^2/\sum w_i(y_{io})^2]^{1/2}$ and the goodness of fit, $\chi^2 = [R_{wp}/R_{exp}]^2$ where $R_{exp} = [(N-P)/\sum w_i y_{io}{}^2]^{1/2}$, y_{io} and y_{ic} are the observed and calculated intensities, w_i is the weighting factor, N is the total number of y_{io} data when the background is refined, and P is the number of adjusted parameters. R_{Bragg} and R_{mag} are the R factors, $\sum|I_{ko} - I_{kc}|/\sum|I_{ko}|$, for nuclear and magnetic peaks, respectively, where I_{ko} and I_{kc} are the observed and calculated integrated intensity. B_{iso} is the isotropic temperature factor. The bond valence sum method was applied to estimate the valence of cations using tabulated parameters[34].

31. Adler, P. *et al.* Structural phase transition in $Sr_2Fe_2O_5$ under high pressure. *J. Solid-State Chem.* **155**, 381–388 (2000).
32. Izumi, F. & Ikeda, T. Rietveld-analysis program RIETAN-98 and its applications to zeolites. *Mater. Sci. Forum* **321–324**, 198–203 (2000).
33. Rodríguez-Carvajal, J. Recent advances in magnetic-structure determination by neutron powder diffraction. *J. Phys. B* **192**, 55–69 (1993).
34. Brown, I. D. & Altermatt, D. Bond-valence parameters obtained from a systematic analysis of the inorganic crystal structure database. *Acta Crystallogr. B* **41**, 244–247 (1985).

LETTERS

Effect of remote sea surface temperature change on tropical cyclone potential intensity

Gabriel A. Vecchi[1] & Brian J. Soden[2]

The response of tropical cyclone activity to global warming is widely debated[1–10]. It is often assumed that warmer sea surface temperatures provide a more favourable environment for the development and intensification of tropical cyclones, but cyclone genesis and intensity are also affected by the vertical thermodynamic properties of the atmosphere[1,10–13]. Here we use climate models and observational reconstructions to explore the relationship between changes in sea surface temperature and tropical cyclone 'potential intensity'—a measure that provides an upper bound on cyclone intensity[10–14] and can also reflect the likelihood of cyclone development[15,16]. We find that changes in local sea surface temperature are inadequate for characterizing even the sign of changes in potential intensity, but that long-term changes in potential intensity are closely related to the regional structure of warming; regions that warm more than the tropical average are characterized by increased potential intensity, and vice versa. We use this relationship to reconstruct changes in potential intensity over the twentieth century from observational reconstructions of sea surface temperature. We find that, even though tropical Atlantic sea surface temperatures are currently at a historical high, Atlantic potential intensity probably peaked in the 1930s and 1950s, and recent values are near the historical average. Our results indicate that—per unit local sea surface temperature change—the response of tropical cyclone activity to natural climate variations, which tend to involve localized changes in sea surface temperature, may be larger than the response to the more uniform patterns of greenhouse-gas-induced warming.

Potential intensity (PI) represents a theoretical upper limit on the intensity of tropical cyclones based on sea surface temperature (SST) and the local vertical thermodynamic structure of the atmosphere[10–12]. With all other factors being equal, a local warming of SST would act to destabilize the overlying atmosphere and increase PI (refs 1,10–13). However, remote SST changes can also influence PI through their influence on upper atmospheric temperatures[1,17]. In the tropical free troposphere, where the Coriolis force is weak, temperature gradients are small and, on timescales longer than a few months, upper tropospheric temperature anomalies are determined by changes in the tropical-mean SST[18]. Thus, local PI in the tropics is influenced by both local and remote SST changes[1,17]. This can be seen on interannual timescales, with Pacific warming leading to increased stability in the tropical Atlantic and contributing to reduced tropical cyclone activity[19], even though Atlantic SSTs are anomalously warm during El Niño.

The effects of local and remote surface warming on PI changes are clearly evident in climate model projections of the twenty-first century performed for the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC-AR4; see Fig. 1). Tropical SST is projected to warm everywhere in response to the increasing greenhouse gases, although the warming is not spatially homogeneous (Fig. 1a). The maximum warming is along the Equator, the Northern Hemisphere warms more than the Southern, and the largest area of warming is in the Indo-Pacific. In the Northern Hemisphere Atlantic, a local minimum in warming covers a broad area, extending from the Caribbean Sea to the northwest coast of Africa. In contrast to the SST, which increases everywhere, the changes in PI are mixed, with regions of both increase and decrease (Fig. 1b). PI is projected to increase in most regions of tropical cyclone activity during the months considered (June–November), except over a broad region in the tropical North Atlantic, where it decreases despite substantial warming. The projections of northern tropical Atlantic PI change show regions of both large increase and moderate decrease. This figure emphasizes that warmer SSTs, by themselves, do not necessarily indicate a more favourable thermodynamic environment for tropical cyclone intensity[1]. This behaviour is highlighted by the contours in Fig. 1b, which show the departure of local SST change from the tropical mean. Notice that decreased PI is associated with regions that warm less than the tropical mean and vice versa. The differing behaviour of PI and SST is also evident in time series of the PI response to global warming (Fig. 2 illustrates the behaviour with one model; Supplementary Figs 1–3 show the 22 IPCC-AR4 models used). Local SSTs increase steadily in the tropics, whereas PI shows large decadal variability, and may increase, decrease or show no change in the long term.

Potential intensity can be computed from atmospheric observational analysis products[20,21] since 1958. The regional structure of PI trends in reanalysis products, and the difference between the two products, can be attributed largely to the regional structure of the SST changes (Supplementary Fig. 4). Because the stabilizing effects of remote SST changes can be estimated by the tropical-mean SST change, regions that warm more (less) than the tropical mean show an increase (decrease) of PI, a behaviour similar to that in model projections of global warming (Fig. 1b).

These results suggest that, when vertical atmospheric profiles are unavailable, the departure of local SST changes from tropical-mean SST changes can be used as a surrogate for PI. Although PI is straightforward to compute, it is dependent on the availability of vertical temperature and humidity data. This is not an issue for model simulations, but it can become problematic for historical reconstructions because observations of vertical profiles of temperature and humidity are only available since the mid-twentieth century and are subject to discontinuities in sampling and data quality. Thus problems may exist with multi-decadal trends computed from data that span large changes in observational practices[12].

Here we make use of the relationship between PI and the local departure from tropical SST change to develop a proxy index for

[1]Geophysical Fluid Dynamics Laboratory, NOAA, Princeton, New Jersey 08542, USA. [2]Rosenstiel School for Marine and Atmospheric Science, University of Miami, Miami, Florida 33149, USA.

changes in PI ($\widetilde{PI}$, see Methods and Supplementary Information). Observational SST reconstructions[22–24] from the 1870s onwards are used to assess the implications of historical SST for long-term changes in PI. Because regional aspects of the long-term changes in SST differ across the reconstructions[25], we explore the $\widetilde{PI}$ changes shown by three different reconstructions (ERSST[22], HadISST[23] and Kaplan[24]; see Supplementary Information) to assess how differences between reconstructions affect $\widetilde{PI}$. In Fig. 3 we show the spatial structure of changes for ref. 22, and those of the other two products in Supplementary Fig. 5.

Between the late nineteenth century and the early twenty-first century, June–November SST has increased throughout the tropics (Fig. 3a, Supplementary Fig. 5). The largest warming has been in the northern and near-equatorial Indian Ocean, and the southern tropical and near-equatorial Atlantic.

In contrast to the SST trends, which are overwhelmingly positive, trends in June–November $\widetilde{PI}$ (Fig. 3b and Supplementary Fig. 5) are mixed, highlighting the spatial heterogeneity of the warming and suggesting that the long-term changes in PI have included regions of both increase and decrease. For example, all products show an increase in $\widetilde{PI}$ in the northern Indian Ocean, a decrease in the western tropical Pacific, and mixed changes in the tropical Atlantic (with regions of both increase and decrease). Thus, even in the presence of warming ocean temperatures over the last century, $\widetilde{PI}$ suggests that the thermodynamic environment in many regions has become less favourable for intensification of tropical cyclones.

Figure 4 shows time series of SST and $\widetilde{PI}$ for the three regions used in Supplementary Fig. 1 and highlighted in Fig. 1a. All three SST data sets indicate substantial warming in the three regions over the twentieth century. In the Atlantic sector, SSTs have been at unprecedented levels since the late 1990s, yet the tropical Atlantic $\widetilde{PI}$ is at near-average levels for that period, and had its highest levels during the middle of the twentieth century (Fig. 4c). The only long-term increase in $\widetilde{PI}$ has been in the Indian Ocean, and recent Pacific $\widetilde{PI}$ has been lower than the long-term mean (the decrease arising abruptly in the 1970s).

The combined influence of local and remote SST changes on $\widetilde{PI}$ can be seen clearly in the Atlantic basin. Atlantic $\widetilde{PI}$ began to decrease in the mid-1950s, even though local SST was not changing substantially ($\widetilde{PI}$ decreases by 0.6–0.7 °C from the 1950s to the 1980s, while local SST decreases by only 0.1–0.2 °C). This reduction in $\widetilde{PI}$ was not dominated by a local SST decrease, but by the rapid warming elsewhere in the tropics (much of it in the Indian Ocean). Until recently, the warming of the Indian Ocean has acted to stabilize the atmosphere elsewhere in the tropics, modulating the $\widetilde{PI}$ changes that would have otherwise occurred. In the most recent decades (over which Atlantic hurricane intensities have increased[2,4,5,7]) $\widetilde{PI}$ has increased substantially in the Atlantic, as the warming of the

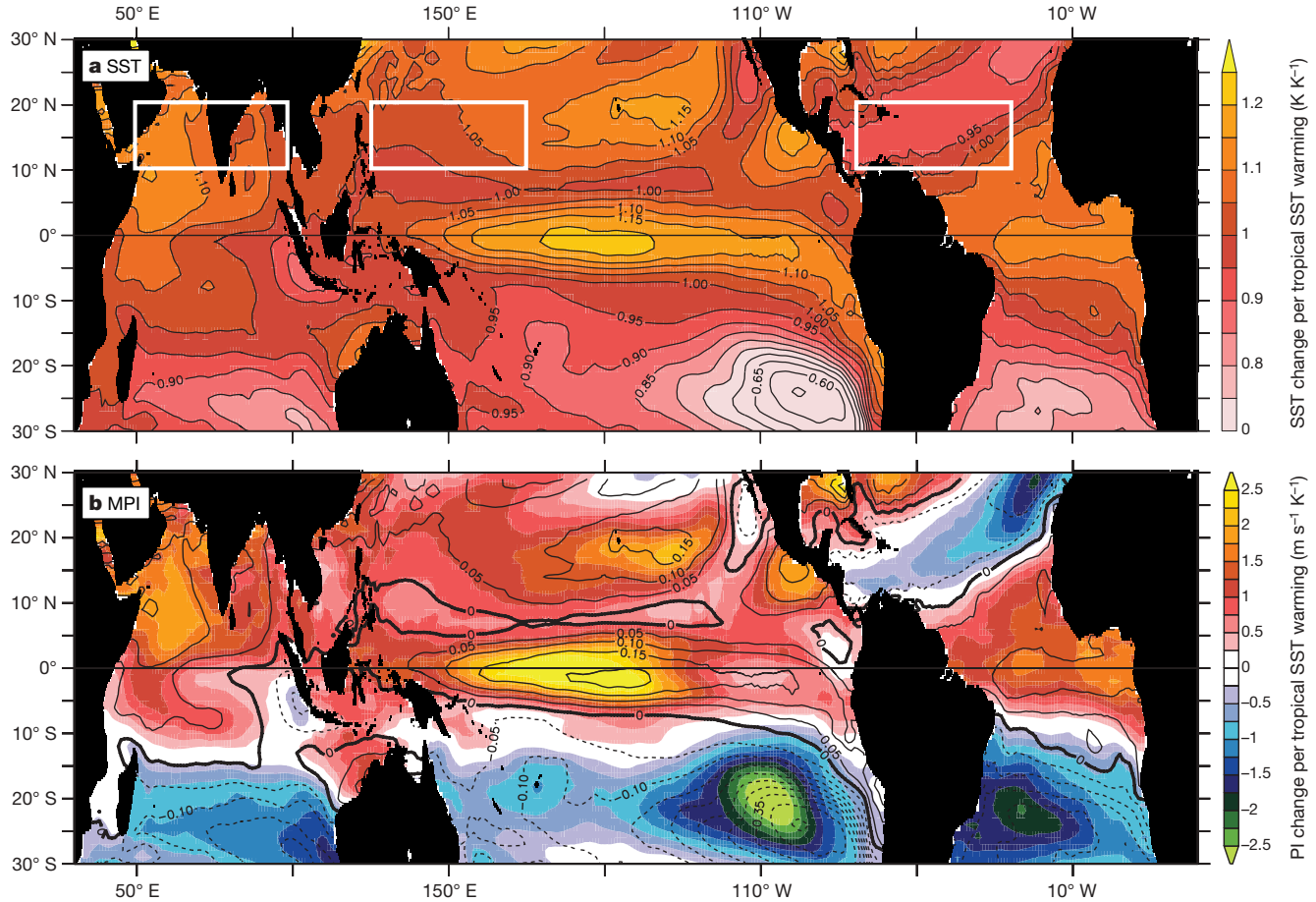


Figure 1 | Spatial structure of model-projected changes in SST and PI for the twenty-first century. The panels show multi-model projections from the IPCC-AR4 emissions scenario A1B (see Methods; this model archive is now known as the World Climate Research Programme Coupled Model Inter-comparison Project 3 database, or CMIP-3). Projections of June–November change per °C tropical warming in **a**, SST (°C °C^{-1}), and **b**, PI (shaded, in m s^{-1} °C^{-1}; ref. 11) and in the normalized departure of the local SST change from the tropical-mean SST change (contoured, in °C °C^{-1}). MPI, maximum PI. The spatial correlation coefficient between SST and PI changes is $r = 0.84$. Changes are normalized by each model's global mean June–November surface air temperature change before averaging. Boxes indicate three regions with time series shown in Fig. 4. This general pattern of surface warming is a robust result of the IPCC-AR4 model projections for the twenty-first century: all of the models show a maximum warming along the Equator[26] and the meridional asymmetry across the Equator, and 16 of those 22 models show the local minimum in the North Atlantic.

Indian Ocean has lessened but warming in the Atlantic has accelerated. Even though Atlantic PI changes since the 1950s have not tracked those of SST, Atlantic tropical cyclone activity has shown a pronounced increase and tracked SST since the 1950s (refs 2,4,10), suggesting that the tropical cyclone activity changes involve factors other than PI as derived in ref. 11, such as vertical wind shear or atmospheric humidity[2,6,9,10,25].

The rapid twentieth -century warming of the Indian Ocean had a different effect on $\tilde{PI}$ locally than it did on the Atlantic. As Indian Ocean SSTs warmed rapidly between the mid-1950s and the 1990s they were tracked by an increase in local $\tilde{PI}$. In recent years, Indian Ocean temperatures have remained relatively steady while the rest of the tropics have continued to warm, and this has led to a decrease in Indian Ocean $\tilde{PI}$.

Although $\tilde{PI}$ is able to capture much of the structure of PI changes, it cannot describe tropical-mean changes in PI. However, regional changes in PI tend to be substantially larger in magnitude than tropical-mean changes, and tropical-mean PI changes can be positive or negative and are poorly constrained by tropical-mean SST changes (see Supplementary Fig. 6). The response of tropical-mean MPI is nominally positive for the models shown in Fig. 1, but is not significantly different from zero given the large inter-model spread; efforts should be undertaken to understand and constrain changes in tropical-mean PI.

These results emphasize the importance of understanding the regional structure of SST changes in response to anthropogenic forcing. For example, the projected increase in Northern Hemisphere June–November PI in response to increased CO_2 (Fig. 1b) is not tied to the 'global' nature of the warming (see Supplementary Fig. 6), but rather to the fact that the model warming shows a meridional asymmetry, being larger in the Northern Hemisphere tropics (Fig. 1a). The prominent decreases in PI in the subsiding regions of the Southern Hemisphere (Fig. 1b) would not directly influence storm intensity (as tropical cyclones do not typically form in those regions), yet the remote influence of the Southern Hemisphere minimum in warming would be to enhance the PI increase in the Northern Hemisphere. The equatorial intensification of tropical ocean warming in model projections for the twenty-first century[26] places the maximum tropical PI increase equatorward of 5° latitude, where planetary vorticity is of small amplitude and thus limits the direct impact of the largest projected SST increases on tropical cyclones; however, the large projected near-equatorial warming also acts to moderate remote PI increases.

A corollary of these results is that localized SST changes are more effective at altering PI than a more uniform temperature change of the same magnitude (see Supplementary Information). This suggests that surface temperature changes driven by well-mixed greenhouse gases, for which local changes in the tropics are dominated by a mean warming (Fig. 1a), will be less effective at modifying PI (per degree local warming) than those driven by internal modes of climate variability, which tend to have much larger spatial gradients in surface warming. For example, given the regional nature of the SST changes associated with the Atlantic Multi-decadal Oscillation[6], the Atlantic Meridional Mode[27], variations of the Atlantic Warm Pool[28] or

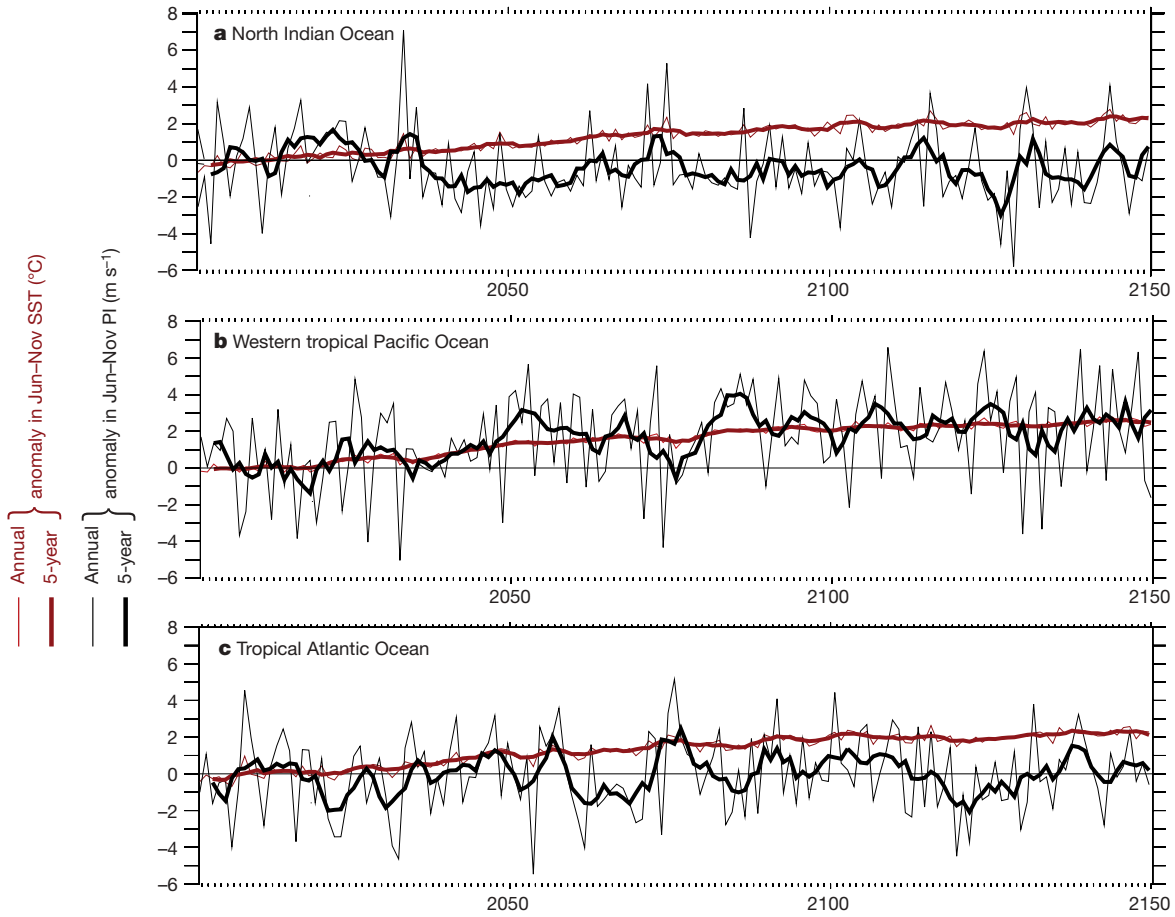


Figure 2 | Time series of June–November change in SST and PI. Changes in SST (red lines, °C) and PI as derived in ref. 11 (black lines, m s^{-1}) are from the GFDL CM2.1 scenario A1B projection in Northern Hemisphere tropical regions (see Fig. 1a). Thin lines show the annual values, thick lines show the 5-yr running mean. Changes calculated from 2001–2020 average. Time series of the changes in SST and PI over these regions for all 22 IPCC-AR4 models can be seen in Supplementary Figs 1–3. Notice that although SST warms steadily in each tropical region, PI shows substantial variability on many timescales, which can overwhelm the long-term change.

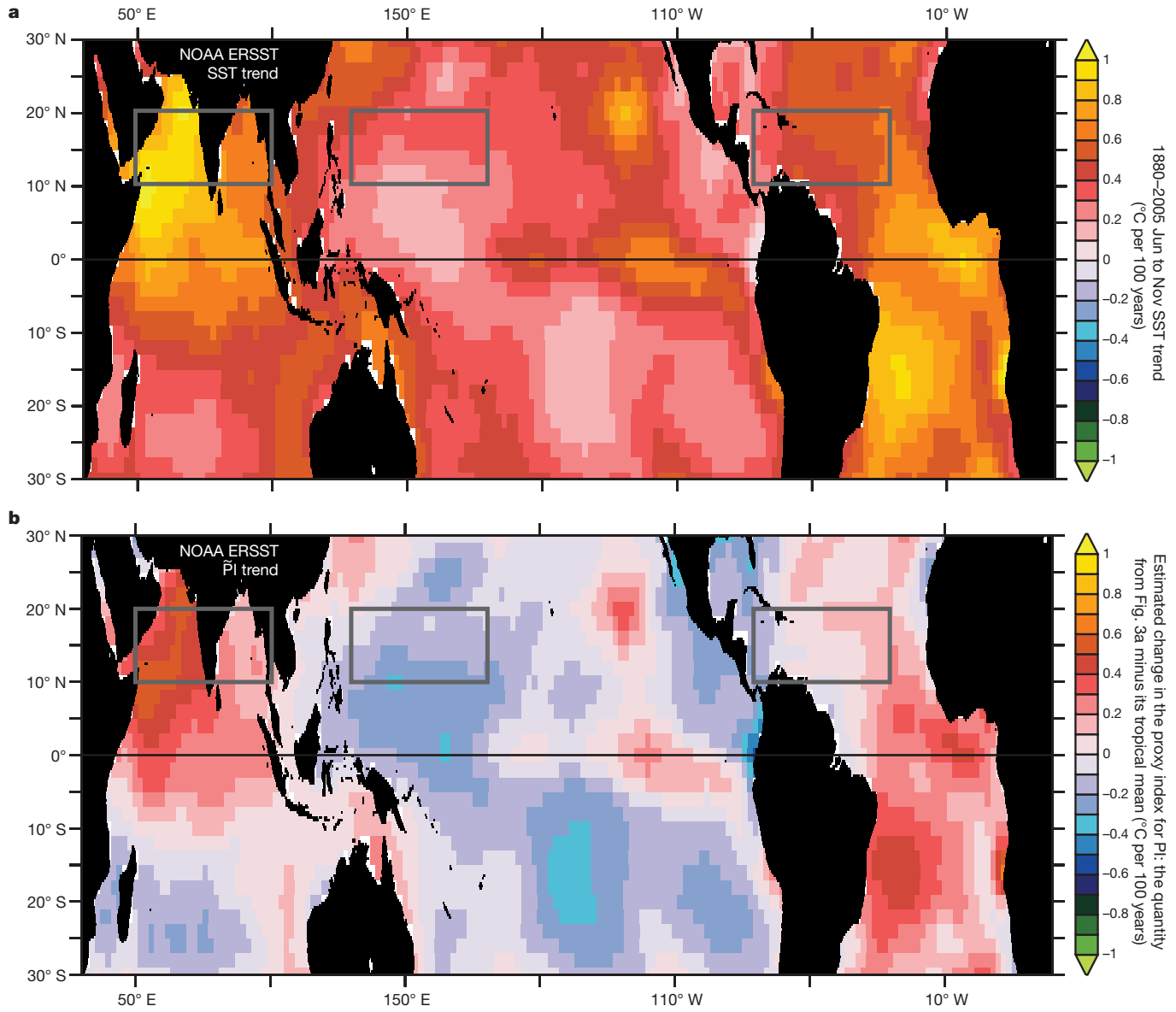


Figure 3 | Century-scale trends in SST and an estimate for PI. a, Linear trend from 1880 to 2006 in June–November SST and **b**, departure of local SST from tropical-mean SST, calculated using the SST reconstructions of ref. 22. See Supplementary Fig. 5 for equivalent figures using refs 24 and 25. Units are °C per century.

differential temperature changes in the Indo-Pacific and Atlantic basins[8], these climate variations should have considerable impact on Atlantic PI in addition to their associated wind shear changes—much as El Niño does[19]. Thus, because of their combined influence of wind shear and PI, these modes of climate variability should have a substantial impact on tropical cyclone activity. More speculatively, given the inhomogeneous character of model[29] and observational[30] estimates of SST anomalies during the Last Glacial Maximum, one may expect that there may have been regions where PI was larger than today, even though the world was considerably colder.

METHODS SUMMARY

We use the potential intensity (PI) derived in ref. 11 to characterize the large-scale thermodynamic environment for tropical cyclone intensification. From model and reanalysis data we compute PI using monthly mean atmospheric temperature, specific humidity, sea level pressure and sea surface temperature (SST) data, and the algorithm available at http://wind.mit.edu/~emanuel/home.html.

We explore changes of SST and PI projected for the twenty-first century using a suite of coupled ocean–atmosphere models forced by emissions scenario A1B (a mid-range emissions scenario, with atmospheric CO_2 stabilization at 720 p.p.m. by year 2100) for the IPCC-AR4. We compute differences between two 20-yr periods: 2001–2020 and 2081–2100. We produce a multi-model ensemble by averaging the response of 22 coupled general circulation models. See online Methods Section for more details. These models do not explicitly resolve the details of tropical cyclones.

Across the reanalyses and the IPCC-AR4 model projections, long-term changes in the difference between local and tropical-mean SST change are strongly correlated with those in local PI ($r \approx 0.8$), and the linear regression fit between the two has a near-zero intercept (see Supplementary table). So we define a proxy index for PI ($\tilde{P}I$) as the difference between local and tropical-mean SST change. Further, the slope of the linear regression fit of PI to $\tilde{P}I$ in June–November is similar in the IPCC-AR4 multi-model ensemble (8.2 m s^{-1} per °C, although it can vary between models; see Supplementary Information) and the reanalyses (8.6 and 8.2 m s^{-1} per °C for refs 21 and 22, respectively), so we use a slope of 8 m s^{-1} to estimate the wind speed equivalent.

By construction, the tropical-mean trend in $\tilde{P}I$ must be 0, whereas tropical-mean PI changes can be slightly non-zero. However, both model projections and historical reanalyses indicate that the tropical-mean multi-decadal changes in PI are typically substantially smaller than the prominent regional changes (see Supplementary Information).

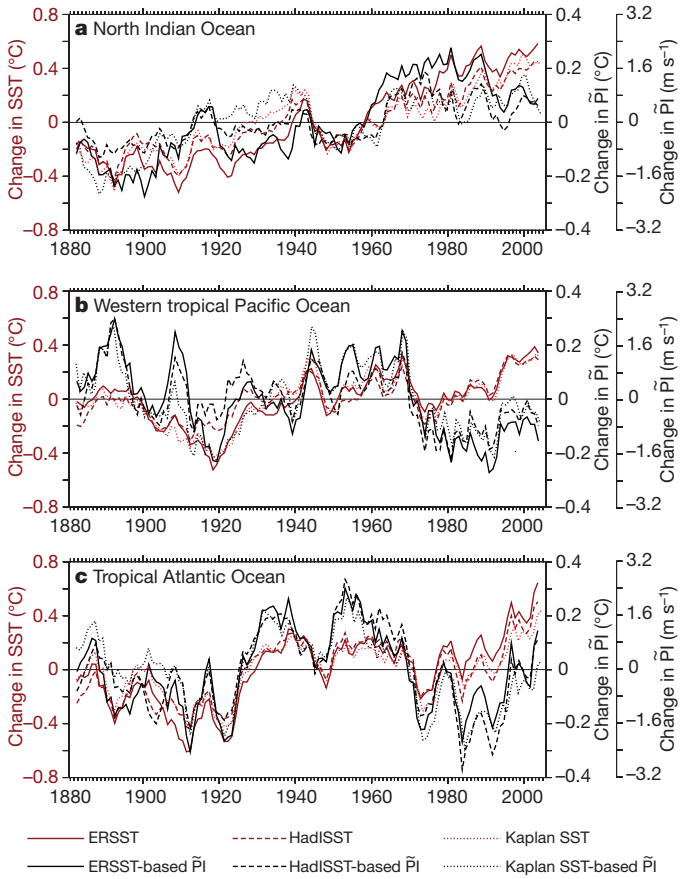


Figure 4 | Anomalies in SST and estimated PI since the late nineteenth century. Regional averages of 5-yr running averaged anomalies in June–November SST (red) and $\tilde{P}I$ (black) based on reconstructions of SST in ref. 23 (solid lines), ref. 24 (dashed lines) and ref. 25 (dotted lines). Units for SST are °C, and those for $\tilde{P}I$ are either °C or m s^{-1} when the regression coefficient between $\tilde{P}I$ and PI from the reanalyses and climate model ensemble-mean is used (~8 m s^{-1} per °C; see Supplementary Information). Anomalies are calculated from the 1880–2006 average.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 23 July; accepted 26 October 2007.

1. Shen, W., Tuleya, R. E. & Ginis, I. A sensitivity study of the thermodynamic environment on GFDL model hurricane intensity: Implications for global warming. *J. Clim.* **13,** 109–121 (2000).
2. Goldenberg, S. B., Landsea, C., Mestas-Nunez, A. M. & Gray, W. M. The recent increase in Atlantic hurricane activity. *Science* **293,** 474–479 (2001).
3. Knutson, T. R. & Tuleya, R. E. Impact of CO_2-induced warming on simulated hurricane intensity and precipitation: Sensitivity to the choice of climate model and convective parameterization. *J. Clim.* **17,** 3477–3495 (2004).
4. Emanuel, K. A. Increasing destructiveness of tropical cyclones over the past 30 years. *Nature* **436,** 686–688 (2005).
5. Webster, P. J., Holland, G. J., Curry, J. A. & Chang, H.-R. Changes in tropical cyclone number, duration and intensity in a warming environment. *Science* **309,** 1844–1846 (2005).
6. Zhang, R. & Delworth, T. L. Impact of Atlantic multidecadal oscillations on India/Sahel rainfall and Atlantic hurricanes. *Geophys. Res. Lett.* **33,** L17712, doi:10.1029/2006GL026267 (2006).
7. Knutson, T. R., Sirutis, J. J., Garner, S. T., Held, I. M. & Tuleya, R. E. Simulation of the recent multi-decadal increase of Atlantic hurricane activity using an 18-km grid regional model. *Bull. Am. Meteorol. Soc.* **88** (10), 1549–1565 (2007).
8. Latif, M., Keenlyside, N. & Bader, J. Tropical sea surface temperature, vertical wind shear, and hurricane development. *Geophys. Res. Lett.* **34,** L01710, doi:10.1029/2006GL027969 (2007).
9. Vecchi, G. A. & Soden, B. J. Increased tropical atlantic wind shear in model projections of global warming. *Geophys. Res. Lett.* **34,** L08702, doi:10.1029/2006GL028905 (2007).
10. Emanuel, K. A. Environmental factors affecting tropical cyclone power dissipation. *J. Clim.* (in the press).
11. Bister, M. & Emanuel, K. A. Dissipative heating and hurricane intensity. *Meteorol. Atmos. Phys.* **65,** 233–240, doi:10.1007/BF01030791 (1998).
12. Bister, M. & Emanuel, K. A. Low frequency variability of tropical cyclone potential intensity. 1. Interannual to interdecadal variability. *J. Geophys. Res.* **107,** 4801, doi:10.1029/2001JD000776 (2002).
13. Holland, G. J. The maximum potential intensity of tropical cyclones. *J. Atmos. Sci.* **54,** 2519–2541 (1997).
14. Emanuel, K. A statistical analysis of tropical cyclone intensity. *Mon. Weath. Rev.* **128,** 1139–1152 (2000).
15. Emanuel, K. A. & Nolan, D. S. Tropical cyclones and the global climate system. In *26th Conf. Hurricanes and Tropical Meteorology (American Meteorological Society, Miami, 2004)*. ⟨ftp://texmex.mit.edu/pub/emanuel/PAPERS/em_nolan_extended_2004.pdf⟩
16. Camargo, S. J., Emanuel, K. A. & Sobel, A. H. Use of genesis potential index to diagnose ENSO effects upon tropical cyclone genesis. *J. Clim.* **20,** 4819–4834 (2007).
17. Elsner, J. B., Tsonis, A. A. & Jagger, T. H. High-frequency variability in hurricane power dissipation and its relationship to global temperature. *Bull. Am. Meteorol. Soc.* **87,** 763–768 (2006).
18. Sobel, A. H., Held, I. M. & Bretherton, C. S. The ENSO signal in tropical tropospheric temperature. *J. Clim.* **15,** 2702–2706 (2002).
19. Tang, B. H. & Neelin, J. D. ENSO Influence on Atlantic hurricanes via tropospheric warming. *Geophys. Res. Lett.* **31,** L24204, doi:10.1029/2004GL021072 (2004).
20. Uppala, S. M. *et al.* The ERA-40 reanalysis. *Q. J. R. Meteorol. Soc.* **131,** 2961–3012 (2005).
21. Kalnay, E. *et al.* The NMC/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **77,** 437–471 (1996).
22. Smith, T. M. & Reynolds, R. W. Extended reconstruction of global sea surface temperatures based on COADS data (1854–1997). *J. Clim.* **16,** 1495–1510 (2003).
23. Rayner, N. A. *et al.* Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.* **108,** doi:10.1029/2002JD002670 (2003).
24. Kaplan, A. *et al.* Analyses of global sea surface temperature 1856–1991. *J. Geophys. Res.* **103,** 18567–18589 (1998).
25. Vecchi, G. A. & Soden, B. J. Global warming and the weakening of the tropical circulation. *J. Clim.* **20,** 4316–4340 (2007).
26. Liu, Z., Vavrus, S., He, F., Wen, N. & Zhong, Y. Rethinking tropical ocean response to global warming: The enhanced equatorial warming. *J. Clim.* **18,** 4684–4700 (2005).
27. Vimont, D. J. & Kossin, J. P. The Atlantic Meridional Mode and hurricane activity. *Geophys. Res. Lett.* **34,** L07709, doi:10.1029/2007GL029683 (2007).
28. Wang, C., Enfield, D. B., Lee, S.-K. & Landsea, C. W. Influences of the Atlantic warm pool on western hemisphere summer rainfall and Atlantic hurricanes. *J. Clim.* **19,** 3011–3028 (2006).
29. Broccoli, A. Tropical cooling at the Last Glacial Maximum: An atmosphere–mixed layer ocean model simulation. *J. Clim.* **13,** 951–976 (2000).
30. CLIMAP Project Members. The last interglacial ocean. *Quat. Res.* **21,** 123–224 (1984).
31. Gualdi, S., Scoccimarro, E., Bellucci, A., Grezio, A., Manzini, E. & Navarra, A. The main features of the 20th century climate as simulated with the SGX coupled GCM. *Claris News* **4,** 7–13 (2006).
32. Gordon, H. B. *et al. The CSIRO Mk3 Climate System Model*. Tech. Report 60 (CSIRO Atmospheric Research, Aspendale, Victoria, 2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge the various modelling groups for providing their data, and PCMDI and the IPCC Data Archive at LLNL/DOE for collecting, archiving and making the data readily available. We thank T. Delworth, K. Dixon, S. Garner, D. E. Harrison, I. Held, A. E. Johansson, T. Knutson, R. Stouffer, A. Wittenberg, S. Ilcane and A. Laperra for discussion, and K. Emanuel for comments. This work was partially supported by NASA and NOAA-OGP.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to G.A.V. (Gabriel.A.Vecchi@noaa.gov).

METHODS

We use the thermodynamic potential intensity for tropical cyclones derived by refs 11 and 12 to characterize the large-scale thermodynamic environment for tropical cyclone intensification. From model and reanalysis data we compute PI using monthly-mean atmospheric temperature, specific humidity, sea level pressure and sear surface temperature (SST) data, and the algorithm available at http://wind.mit.edu/~emanuel/home.html.

To explore twenty-first-century projected changes of SST and PI, we use a suite of coupled ocean–atmosphere models forced by emissions scenario A1B (a mid-range emissions scenario, with atmospheric CO_2 stabilization at 720 p.p.m. by year 2100) for IPCC-AR4. We calculate between two 20-yr periods: 2001–2020 and 2081–2100. We produce a multi-model ensemble by averaging the response of 22 coupled general circulation models (the 20 models with three-dimensional atmospheric data used by ref. 9 and described in Table 1 of ref. 26, along with the INGV (ref. 31) and CSIRO-Mk3.5 (ref. 32) models that became available more recently in the IPCC-AR4 database). We also look at two additional emissions scenarios (A2 and B1, with atmospheric CO_2 levels at 2100 of 855 p.p.m. and 550 p.p.m., respectively) to explore the relationship between tropical-mean changes in PI and SST. These models are generally of relatively coarse resolution and do not explicitly resolve tropical cyclone dynamics, but they are able to represent aspects of large-scale climate conditions that are relevant to tropical cyclone genesis and intensification, such as the thermodynamic structure of the atmosphere.

To examine observed changes in PI and SST relative to the late 1950s we use data from the European Centre for Medium Range Weather Forecasting 40-year Atmospheric Reanalysis (ERA40; ref. 20) and from the US National Center for Environmental Prediction (NCEP) Atmospheric Reanalysis (ref. 21). For each reanalysis product long-term changes are computed as linear least-squares trends over the period 1958–2002. As these two model-based atmospheric observational reanalyses are affected by data discontinuities and a changing observing system, it has been argued that these changes affect the validity of the PI values that are calculated from these analyses[12]. Nevertheless, we explore the relationship between changes in SST and PI in these products, to find an 'observational' analogue for the projections of future climate change of the IPCC-AR4 models. It is possible that the overall tendency for a tropical-mean decrease in PI in these products is due to inhomogeneity in observing practices[12], yet the tropical-mean changes in PI are small relative to the regional changes—which are the focus of this study (see Section IV of the Supplementary Information).

We calculate long-term changes in historical SST over the period 1880–2005, using statistical reconstructions of instrumental data, which take historical observed SST data and use statistical methods to reconstruct global gridded data sets. Because the long-term trends in different SST reconstructions can differ[25], we use three historical reconstruction products: the US National Oceanic and Atmospheric Administration Extended Reconstruction of SST (NOAA-ERSST; ref. 22); the UK Meteorological Office Hadley Centre Interpolated SST product (HadISST; ref. 23); and the Columbia University Lamont-Doherty Earth Observatory's Historical SST Recontruction (Kaplan SST; ref. 24).

These three SST products differ in their analysis procedure, with different statistical reconstruction methods used to estimate the value of SST in regions with no data. Also, the methods differ in the corrections applied to the data to account for changes in observing practices (such as going from 'bucket temperature measurements', where the temperature of water taken using buckets was measured, to 'ship intake' temperatures, where the temperature of the water in the engine room intake was measured, to 'hull sensor' measurements, where measurements of temperature are made using sensors mounted on the hulls of ships), with HadISST and Kaplan sharing a correction algorithm that differs from that of ERSST. These data sets also use slightly different data sources (for example, the NOAA product uses only *in situ* measurements over the whole record, whereas the source data for Kaplan and HadISST include satellite-derived SST starting in the early 1980s; the source data for Kaplan and HadISST include additional *in situ* observations from the Met Office archive not present in the NOAA product). Although the overall tendency for SST warming is robust, there are discrepancies in the spatial structure of the changes in all three tropical basins. Until the disagreement between the various SST records is resolved, we believe it prudent to examine all three and view their differences as an estimate of the uncertainty in SST. The true uncertainty in SST is larger than the discrepancy between these SST products, as they share many common elements, so their differences are a lower-bound estimate in uncertainty. Also, and almost self-evidently, the further one extends the record back in time, the larger the uncertainty. Nonetheless, these observationally based estimates of SST allow us to estimate the changes in conditions in the tropics since the late nineteenth century.

LETTERS

Dynamics of Mid-Palaeocene North Atlantic rifting linked with European intra-plate deformations

Søren B. Nielsen[1], Randell Stephenson[2] & Erik Thomsen[1]

The process of continental break-up provides a large-scale experiment that can be used to test causal relations between plate tectonics and the dynamics of the Earth's deep mantle[1,2]. Detailed diagnostic information on the timing and dynamics of such events, which are not resolved by plate kinematic reconstructions, can be obtained from the response of the interior of adjacent continental plates to stress changes generated by plate boundary processes. Here we demonstrate a causal relationship between North Atlantic continental rifting at ~62 Myr ago and an abrupt change of the intra-plate deformation style in the adjacent European continent. The rifting involved a left-lateral displacement between the North American-Greenland plate and Eurasia, which initiated the observed pause in the relative convergence of Europe and Africa[3]. The associated stress change in the European continent was significant and explains the sudden termination of a ~20-Myr-long contractional intra-plate deformation within Europe[4], during the late Cretaceous period to the earliest Palaeocene epoch, which was replaced by low-amplitude intra-plate stress-relaxation features[5]. The pre-rupture tectonic stress was large enough to have been responsible for precipitating continental break-up, so there is no need to invoke a thermal mantle plume as a driving mechanism. The model explains the simultaneous timing of several diverse geological events, and shows how the intra-continental stratigraphic record can reveal the timing and dynamics of stress changes, which cannot be resolved by reconstructions based only on plate kinematics.

Intra-plate basin inversion structures in Europe (Fig. 1) formed initially by transverse shortening and erosion of the central parts of Palaeozoic and Mesozoic era sediment-filled rifts and troughs in response to compressional pulses during the Late Cretaceous, particularly the Campanian and Maastrichtian ages[4]. The shortening produced an internal lithospheric load, as uplifted and eroded lighter sediments were replaced by more compacted sediments and crystalline crust[5]. The presence of these loads and their longevity are corroborated by positive Bouguer gravity signatures[6] along the inversion zones and by the occurrence of flexurally controlled asymmetric primary marginal troughs flanking most of the European inversion structures[5]. The depths and widths of the flexural troughs reflect the magnitudes of the loads and the apparent elastic thickness of the lithosphere, which is of the order of 5–10 km in the basin settings of the European inversion structures[5].

According to the thin elastic plate model (a widely adopted proxy for the effects of low-stress loading of more complex lithospheric rheologies), such flexures provide sensitive barometers of changes in the in-plane tectonic stress[7]. Compression perpendicular to the strike of the structure deepens the flexure, while extension shallows it, and such effects can be preserved by the sedimentary record. Thus, a sudden release of in-plane compression has been invoked to explain the change of deformation style in the evolution of European inversion structures in the mid-Palaeocene (beginning in the Late Danian age, ~62 Myr ago) from one of compressional shortening to one of

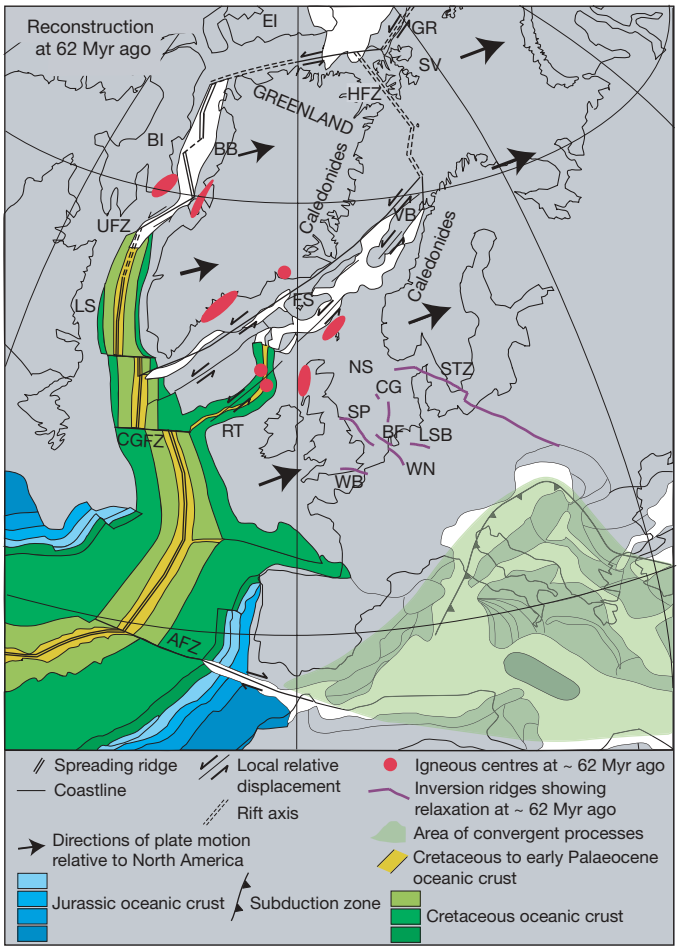


Figure 1 | Regional geological reconstruction[27] at ~62 Myr with present-day coast lines. For modelling purposes, the complex zone of north–south convergence inferred between Europe and Africa is taken to be a single discrete plate boundary. AFZ, Azores fracture zone; BB, Baffin Bay; BF, Broad Fourteens basin; BI, Baffin Island; CG, Central graben; CGFZ, Charlie–Gibbs fracture zone; EI, Ellesmere Island; FS, Faeroe–Shetland trough; GR, future Gakkel rift; HFZ, Hornsund fault zone; LS, Labrador Sea; LSB, Lower Saxony basin; NS, North Sea basin; VB, Vøring basin; RT, Rockall trough; SP, Sole Pit High; SV, Svalbard; UFZ, Ungava fault zone; WB, Weald–Boulonnais area; WN, West Netherlands basin.

[1]Department of Earth Science, University of Aarhus, Høegh-Guldbergsgade 2, DK-8000 Aarhus C, Denmark. [2]Faculty of Earth and Life Sciences, Vrije Universiteit, De Boelelaan 1085, 1081 HV Amsterdam, The Netherlands.

non-ruptural doming of a wider area[5]. This plate-wide stress change has also emerged from regional microtectonic fault studies[8], but is most accurately dated in the flexural trough along the Sorgenfrei–Tornquist Zone (STZ) of the eastern North Sea area, where the Palaeocene stratigraphy is known in detail (Figs 1, 2). The early Danian deposits (nannoplankton Palaeogene NP1–early NP4; see Fig. 2) record a deepening in the direction of the inversion axis, indicating that compressional shortening was still occurring at this time. The late Danian and Selandian (middle NP4–NP7) depositional centre (depocentre), however, occurs in a more distal position that is consistent with an upward flexural doming of the central inversion ridge at the onset of the Late Danian, and an associated flexural downwarp, creating a secondary marginal trough[5]. The amplitude of the flexure is of the order of 10^2m.

Similar but less stringent timing constraints can be derived from other European inversion zones. The Weald–Boulonnais area, for example, is flanked by Palaeocene depocentres. The northernmost depocentre, which was initially in a continental setting that prevented sedimentation before flooding, occurred slightly before the Selandian–Thanetian boundary (~58.5 Myr ago)[5]. The onset of flexural relaxation of inversion structures in the Netherlands is constrained to be during the Middle Palaeocene (early Selandian, ~61 Myr ago), marked by the occurrence of reworked late Cretaceous nannofossils in Middle Palaeocene deposits, derived from erosion of the late Cretaceous inversion ridges[5].

Neither a drop in eustatic sea level at ~62 Myr ago nor differential compaction effects can explain the plate-wide synchronicity of the mid-Palaeocene shifts in depocentres. In the STZ, for example, the lower to middle Danian deposits (NP1–early NP4, Fig. 2) are thin or missing where the (post-62 Myr ago, middle NP4–NP6) secondary marginal trough is thickest. Furthermore, the Danian sediments are entirely autochthonous and biogenic, without any indication of reworking, which would be expected if a sea level drop had been involved.

Thus, a fundamental change in the intra-plate stress field of Europe at ~62 Myr ago is an appropriate and convincing explanation for the intra-continental sedimentary record[5] and the regional history of fault patterns[8]. This change reduced the compressional component perpendicular to the strike of the inversion structures. Here, we analyse how this change might be related to the following two stress-generating, plate-tectonic events affecting the Arctic, the North Atlantic and the European and African continents.

(1) The impingement of a major thermal mantle plume on the base of the North Atlantic lithosphere. At ~62 Myr ago, there was an almost simultaneous outbreak of volcanism on Baffin Island[9], East and West Greenland[10], in the Hebridean igneous province of northwest Scotland and environs[11] and in the Rockall trough[12]. This magmatic episode predated the eruption of the voluminous flood basalts at ~56 Myr ago[10] along the North Atlantic spreading ridge. The synchronicity of volcanism over a large area has typically been taken to mark the rapid spreading of the plume head at the base of the lithosphere. Such an event would cause isostatic uplift and a change in in-plane stress that would affect the flexural equilibrium of the surrounding lithosphere.

(2) Cessation of north–south convergence between Africa and Europe. Plate kinematic reconstructions strongly suggest that the north–south convergence of Africa and Europe ceased for a period of ~10 Myr during the Palaeocene[3]. The dynamic meaning of this convergence break is not known. It may represent a change in the mode of Alpine collision from the subduction of the intervening ocean (Tethys) exerting compression on Europe, to the subduction being replaced by increasing compression in the continental plates[13,14].

We quantified the effects of the two mid-Palaeocene stress-changing tectonic events on the flexural state of the European lithosphere by calculating stress propagation on an elastic spherical shell and using the flexure equations of an elastic plate. The effects of a mantle plume are simulated by using the gradient of its contribution to lithospheric potential energy as body force in the stress equilibrium equations. The generalized effect of Africa pushing on the Eurasian lithosphere is obtained by displacing northern Africa

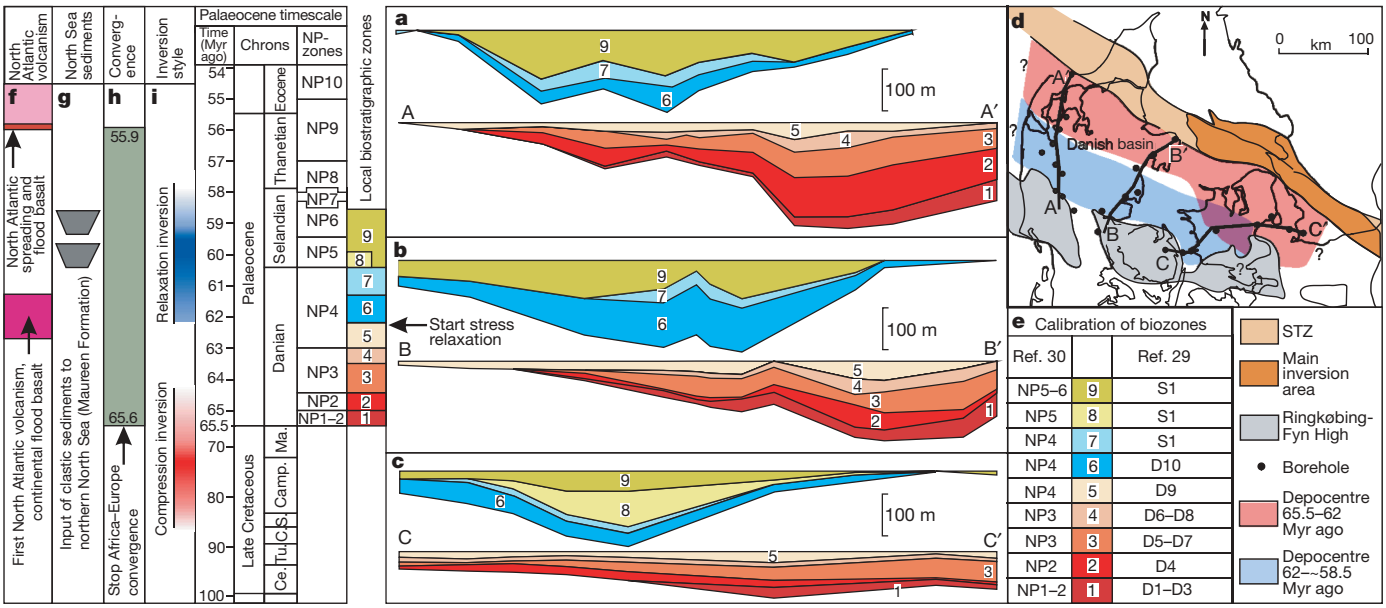


Figure 2 | Correlation of Palaeocene[28] geological events. a–c, Sections across the marginal trough of the STZ showing the distribution of lower and middle Palaeocene (Danian, Selandian) deposits subdivided into nine calcareous nannoplankton (biostratigraphic) zones, each plotted using two datum levels. An abrupt shift in the depocentre occurred at ~62 Myr ago (between nannoplankton zones 5 and 6), reflecting a change from compressional to relaxational inversion style. **d**, Location of sections in **a–c** and the shift in depocentre during the Palaeocene. ?, no borehole control. **e**, Correlation between the nannoplankton zonation scheme used in this paper and the local North Sea zonation scheme[29], and the global NP-zonation scheme defined by Martini[30]. **f**, Onset of North Atlantic volcanism and ridge spreading between Greenland and Norway. **g**, Input of clastic sediments into the northern North Sea (Maureen Formation) coinciding with the Danian–Selandian boundary. **h**, Temporary stop in convergence between Africa and Europe bracketed by datings of 65.6 Myr and 55.9 Myr ago[3]. **i**, Timing and shift in intra-plate (inversion) style in the STZ and other inverted basins in the North Sea Basin. Ce., Cenomanian; Tu., Turonian; C., Coniacian; S., Santonian; Camp., Campanian; Ma, Maastrichtian.

towards the north, while keeping the equatorial region to the south a free boundary (see Supplementary Information). Other in-plane lithospheric stress systems also modify the deflections (Fig. 3a), but do not change rapidly at ~62 Myr ago (and therefore do not contribute to changing the flexural equilibrium) and are not considered further. These include ridge push from the central Atlantic Ocean, ridge push from the incipient accretionary plate boundary in the Labrador Sea[15], stresses from the topography of old mountain ranges like the Caledonides, and other density-related lithospheric loads.

We found that even a large mantle plume (80 km thick at the centre with a gaussian half-width of 1,000 km, producing 1,400 m of surface uplift at the centre) at the base of the Greenland and North Atlantic lithosphere produces only minimal stress effects in the European plate. The far-field compression from the plume (about -0.75×10^{12} N m^{-1}) is aligned with the strike of the structures and has only negligible flexural effect, while a small extension (about 0.75×10^{12} N m^{-1}) perpendicular to their strike causes a minor (of the order of 10^{1} m) flexural uplift (Fig. 3b). We conclude from this that a plume is not likely to be responsible for producing the well-constrained changes of style of the European inversion structures in the mid-Palaeocene.

The north–south convergence of Europe and Africa in the Late Cretaceous and early Palaeocene in our model contributes a north–south oriented compressional stress component, which depresses (deepens) the European inversion structures (Fig. 3c) because the stress is large and because the structures are favourably oriented. This is in keeping with the style of development of these structures at this time[4]. The magnitude of compression was adjusted to $3–4 \times 10^{12}$ N m^{-1}, comparable to a ridge push. This could be the main driving mechanism for Africa, which, apart from in the north, was surrounded by spreading ridges at this time. If the suspension of convergence were associated with a further increase in compression[13,14], the flexural deepening would be enhanced according to our model, conflicting with the observed domal flexural uplift of the inversion zones[5]. In contrast, the observed basin inversion style[5] and stress change[8] requires that the convergence break be associated with a relaxation of the convergence-induced stress state in Europe (that is, of the Eurasian plate).

We suggest that this stress relaxation occurred through left-lateral displacements along a fracture system through the North Atlantic and along the (future) Gakkel ridge of the Arctic Ocean, and was eventually involved in, and relaxed by, Arctic Eurasian tectonic processes on the Siberian continental margin[16]. The magnitude of left-lateral displacements on the fracture system is determined by continent-scale relaxation of the elastic strain state of Eurasia and the North American–Greenland plate, which was created and maintained during the Late Cretaceous and earliest Palaeocene Atlantic opening and Africa–Eurasia interactions. This further implies initiation of mid-Palaeocene (~62 Myr ago) separation between Greenland and Eurasia by extension on the Hornsund fault zone and its south-eastern prolongation (Fig. 1). This is documented by early Palaeocene (Danian) basin initiation in Svalbard[17,18] and the occurrence of strongly stretched continental crust along the western margin of Svalbard[19]. The changing geometry along this transfer rift set the scene for the strikingly different structural developments during the latest Palaeocene-earliest Eocene transition to dextral strike slip[18,19], with the onset of spreading in the North Atlantic and the Arctic Ocean.

The timing of the Africa–Europe convergence break to between 65.6 Myr and 55.9 Myr ago is based on sparse ocean magnetic anomaly age data and interpolation of Euler poles of a finite precision[3]. This timing is poorly constrained compared to the ~300,000-year resolution of the continental stratigraphic record at this time (Fig. 2). We therefore suggest that the occurrence of the stratigraphically dated relaxation flexures at ~62 Myr ago marks the onset of the punctuation in relative convergence, which then acquires a dynamic interpretation in terms of the 'escape' of Eurasia from the impinging African continent.

Evidence of mid-Palaeocene left-lateral displacements on fault systems along the proto-North Atlantic exists. Left-lateral displacement on a fracture reaching from the Vøring basin offshore of northern Norway and passing through the Faeroe–Shetland basin and the Rockall trough to terminate in the Charlie–Gibbs fracture zone of the Atlantic Ocean (Fig. 1) has been inferred for the latest Danian–Selandian times[20]. This fracture pathway tracks the trend of the Palaeozoic and Mesozoic rift system that developed during the region's protracted post-Caledonian rifting history. Left-lateral faulting activity with displacements in the 1–2 km range in the north-eastern Vøring basin began in latest Maastrichtian/earliest Palaeocene times and intensified before the onset of ocean spreading at ~56 Myr (refs 21 and 22). Rifting in the Faeroe-Shetland basin, possibly with a strike-slip component, occurred in the early Palaeocene and terminated at ~59.5 Myr before the onset of a brief compressional phase at ~56 Myr (ref. 23), possibly related to the onset of North Atlantic spreading. Furthermore, northeast–southwest extension in the Hebridean igneous province has been inferred from the orientation of dykes in the area[24]. The timing of rifting in the Rockall trough is at present equivocal because of a lack

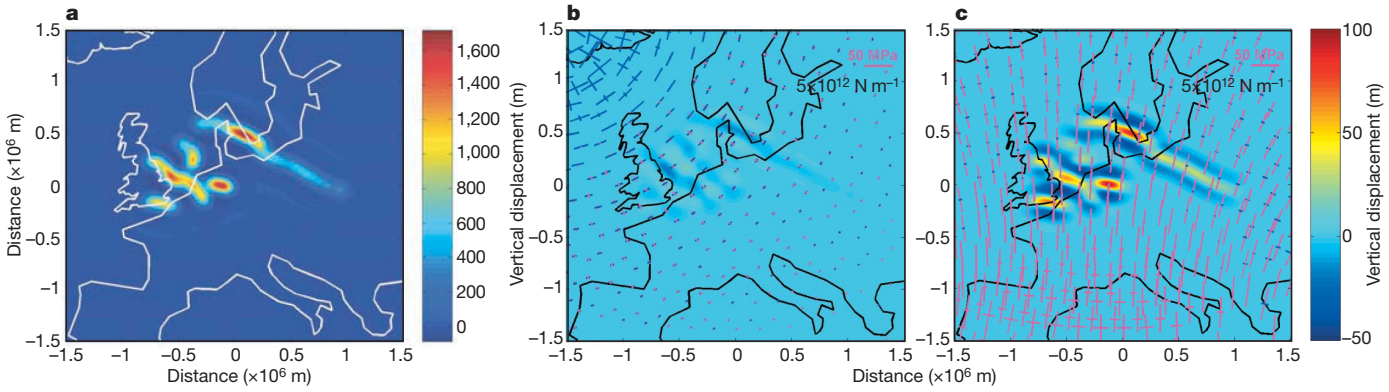


Figure 3 | Model results. Colour bars show vertical displacement. The right colour bar is shared by **b** and **c**. **a**, Vertical displacement of the flexural surface caused by the load of inversion ridges. The deflections host the primary marginal troughs and the inversion ridges (not shown) in their axes. **b**, Flexural effect of swell push from a proto-Icelandic plume. Plume stress could have caused a small flexural uplift because of a small component of extension perpendicular to the strike of the structures. **c**, Flexural over-deepening caused by convergent interaction of Africa and Europe. Flexural uplift of the same magnitude occurred when stresses relaxed because of left-lateral relative motion of North America–Greenland and Europe starting at ~62 Myr ago. Crosses in **b** and **c** represent principal stress directions and magnitudes taken from the spherical stress model (Supplementary Information) and projected onto a flat map for flexural calculations. Purple represents compression (scale bar) and blue represents extension. The flexural elastic plate thickness is 7 km.

of well control and poor seismic resolution below basalts, but it is believed to be mainly of Late Cretaceous age.

In this context, the acceleration of extension and sea floor spreading in Labrador Sea and Baffin Bay at ~62 Myr ago[15], and the associated anti-clockwise, rigid, rotation of Greenland away from Canada, are large-scale expressions of the left-lateral release between Eurasia and the North American–Greenland craton. That this dominates over any dextral displacement of the Greenland block relative to Europe implied by opening in the Labrador Sea–Baffin Bay spreading system suggests that the latter was induced by the European escape rather than driving it. We take the unusual density of pseudotachylyte breccias (dated at 62.9 ± 4.5 Myr ago) along normal and strike-slip faults in East Greenland to indicate the rapidity of this tectonic event[25].

Our model brings a number of diverse geological observations into a unifying framework, including North Atlantic rupture and African–European convergence. The occurrence of mid-Palaeocene stress relaxation in the European continent is well documented by the observation of macrotectonic relaxation flexures[5] and from microtectonic fault analysis[8]. Stress relaxation was simultaneous with the first outbreaks of North Atlantic volcanism ~62 Myr ago. This, together with structural evidence from the North Atlantic and the Svalbard area, and the temporary cessation of African–Eurasian convergence, have led us to conclude that the stress relaxation marked the onset of a plate-scale left-lateral translation between the North American–Greenland and Eurasian plates, which relieved continent-scale elastic strain. This implies the existence of a significant pre-rupture in-plane stress, which could have been responsible for the onset of left-lateral rupture without a causative convective event in the mantle[2]. Continental rupture driven solely by plume uplift would have produced a poloidal[1] (extension/subduction) lithospheric velocity field rather than the toroidal (transform/spin) field inferred here. This further suggests that rapid rifting in the area of the North Atlantic Caledonide suture might have triggered the observed continental-style volcanism[26] starting ~62 Myr ago, thus precluding the requirement of a rapidly spreading plume head under the North Atlantic lithosphere to explain the simultaneous widespread outbreaks of volcanism.

Received 13 March; accepted 10 October 2007.

1. Lithgow-Bertelloni, C. & Richards, M. A. The dynamics of Cenozoic and Mesozoic plate motions. *Rev. Geophys.* **36,** 27–43 (1998).
2. Anderson, D. L. Top-down tectonics? *Science* **293,** 2016–2018 (2001).
3. Rosenbaum, G., Lister, G. S. & Duboz, C. Relative motions of Africa, Iberia and Europe during Alpine orogeny. *Tectonophysics* **359,** 117–129 (2002).
4. Ziegler, P. A. *Geological Atlas of Western and Central Europe* 1–239 (Shell International Petroleum, Den Haag, 1990).
5. Nielsen, S. B., Thomsen, E., Hansen, D. L. & Clausen, O. R. Plate-wide stress relaxation explains European Palaeocene basin inversions. *Nature* **435,** 195–198 (2005).
6. Wybraniec, S. *et al.* New map compiled of Europe's gravity field. *Eos* 79 (37), 437–442 (1998.).
7. Cloetingh, S., McQueen, H. & Lambeck, K. On a tectonic mechanism for regional sea level variations. *Earth Planet. Sci. Lett.* **75,** 157–166 (1985).
8. Vandycke, S. Palaeostress records in Cretaceous formations in NW Europe: extensional and strike–slip events in relationships with Cretaceous–Tertiary inversion tectonics. *Tectonophysics* **357,** 119–136 (2002).
9. Kent, A. J. R. *et al.* Mantle heterogeneity during the formation of the North Atlantic Igneous Province: Constraints from trace element and Sr-Nd-Os-O isotope systematics of Baffin Island picrites. *Geochem. Geophys. Geosyst.* **5,** Q11004, doi:10.1029/2004GC000743 (2004).
10. Tegner, C. *et al.* ^{40}Ar-^{39}Ar geochronology of Tertiary mafic intrusions along the East Greenland rifted margin: relation to flood basalts and the Iceland hotspot track. *Earth Planet. Sci. Lett.* **156,** 75–88 (1998).
11. Ritchie, J. D., Gatliff, R. W. & Richards, P. C. in *Petroleum Geology of Northwest Europe: Proc. 5th Conf.* (eds Fleet, A. J. & Boldy, S. A. R.) 573–584 (Geological Society, London, 1999).
12. O'Connor, J. M., Stoffers, P., Wijbrans, J. R., Shannon, P. M. & Morrissey, T. Evidence from episodic seamount volcanism for pulsing of the Iceland plume in the past 70 Myr. *Nature* **408,** 954–958 (2000).
13. Dèzes, P., Schmid, S. M. & Ziegler, P. Evolution of the European Cenozoic rift system: interaction of the Alpine and Pyrenean orogens with their foreland lithosphere. *Tectonophysics* **389,** 1–33 (2004).
14. Ziegler, P. A., Cloetingh, S. & van Wees, J.-D. Dynamics of intra-plate compressional deformation: the Alpine foreland and other examples. *Tectonophysics* **252,** 7–59 (1995).
15. Roest, W. R. & Srivastava, S. P. Sea-floor spreading in the Labrador Sea—a new reconstruction. *Geology* **17,** 1000–1003 (1989).
16. Franke, D., Hintz, K. & Oncken, O. The Laptev Sea rift. *Mar. Petrol. Geol.* **18,** 1083–1127 (2001).
17. Nagy, J. Delta-influenced foraminiferal facies and sequence stratigraphy of Paleocene deposits in Spitsbergen. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **222,** 161–179 (2005).
18. Saalmann, K. & Thiedig, F. Thrust tectonics on Broggerhalvoya and their relationship to the Tertiary West Spitsbergen fold-and-thrust belt. *Geol. Mag.* **139,** 47–72 (2002).
19. Faleide, J. I., Gudlaugsson, S. T., Eldholm, O., Myhre, A. M. & Jackson, H. R. Deep seismic transects across the sheared western Barents Sea-Svalbard continental margin. *Tectonophysics* **189,** 73–89 (1991).
20. Harrison, J. C. *et al.* Correlation of Cenozoic sequences of the Canadian Arctic region and Greenland: implications for the tectonic history of northern North America. *Bull. Can. Petrol. Geol.* **47,** 223–254 (1999).
21. Mogensen, T. E., Nyby, R., Karpuz, R. & Haremo, P. *Late Cretaceous and Tertiary Structural Evolution of the Northeastern Part of the Vøring Basin, Norwegian Sea* 379–396 (Spec. Publ. 167, Geological Society, London, 2000).
22. Imber, J. *et al.* Early Tertiary sinistral transpression and fault reactivation in the western Voring Basin, Norwegian Sea: Implications for hydrocarbon chloration and pre-breakup deformation in ocean margin basins. *Bull. Am. Assoc. Petrol. Geol.* **89,** 1043–1069 (2005).
23. Dean, K., McLachlan, K. & Chambers, A. in *Petroleum Geology of Northwest Europe: Proc. 5th Conf.* (eds Fleet, A. J. & Boldy, S. A. R.) 533–544 (Geological Society, London, 1999).
24. England, R. W. *The Early Tertiary Stress Regime in NW Britain: Evidence From the Patterns of Volcanic Activity* 381–389 (Spec. Publ. 39, Geological Society, London, 1988).
25. Karson, J. A., Brooks, C. K., Storey, M. & Pringle, M. S. Tertiary faulting and pseudotachylytes in the East Greenland volcanic rifted margin: seismogenic faulting during magmatic construction. *Geology* **26,** 39–42 (1998).
26. Christiansen, R. L., Foulger, G. R. & Evans, J. R. Upper-mantle origin of the Yellowstone hotspot. *Geol. Soc. Am. Bull.* **114,** 1245–1256 (2002).
27. Schettino, A. & Scotese, C. R. New Internet software aids paleomagnetic analysis and plate tectonic reconstructions. *Eos* **82,** 530–536 (2001).
28. Luterbacher, H. P. *et al.* in *A Geologic Time Scale* (eds Gradstein, F. M., Ogg, J. G. & Smith, A. G.) 384–408 (Cambridge Univ. Press, Cambridge, 2004).
29. Perch-Nielsen, K. in *Proc. Cretaceous–Tertiary Boundary Events Symp.* (eds Birkelund, T. & Bromley, R. G.) Vol. 1 115–135 (Univ. of Copenhagen, Copenhagen, 1979).
30. Martini, E. in *Proc. II Planktonic Conf. (Roma 1970)* (ed. Fainacci, A.) Vol. 2 739–785 (Tecnoscienza, Rome, 1971).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was initiated during a visiting fellowship for R. Stephenson at the Department of Earth Sciences, Aarhus, and completed during project COLD, supported by the Danish Natural Science Research Council.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to S.B.N. (sbn@geo.au.dk).

LETTERS

Fetal load and the evolution of lumbar lordosis in bipedal hominins

Katherine K. Whitcome[1], Liza J. Shapiro[2] & Daniel E. Lieberman[1]

As predicted by Darwin[1], bipedal posture and locomotion are key distinguishing features of the earliest known hominins[2,3]. Hominin axial skeletons show many derived adaptations for bipedalism, including an elongated lumbar region, both in the number of vertebrae and their lengths, as well as a marked posterior concavity of wedged lumbar vertebrae, known as a lordosis[4–6]. The lordosis stabilizes the upper body over the lower limbs in bipeds by positioning the trunk's centre of mass (COM) above the hips. However, bipedalism poses a unique challenge to pregnant females because the changing body shape and the extra mass associated with pregnancy shift the trunk's COM anterior to the hips. Here we show that human females have evolved a derived curvature and reinforcement of the lumbar vertebrae to compensate for this bipedal obstetric load. Similarly dimorphic morphologies in fossil vertebrae of *Australopithecus* suggest that this adaptation to fetal load preceded the evolution of *Homo*.

Until recently, hominin females spent most of their adult lives either pregnant or lactating[7]. Pregnancy augments the mass of the human female abdomen by as much as 31% (6.8 kg)[8], translating the position of the maternal COM forward and increasing the torque exerted by the upper body around the hip joints. Although this shift in mass does not disrupt postural stability in quadrupeds (Fig. 1a, b), it uniquely destabilizes bipeds whose supporting joints and two-footed support base lie solely under the hips (Fig. 1c, d). Such gravid instability can be counteracted by muscles, but sustained recruitment risks muscle fatigue and increases the likelihood of spinal injury[9].

Pregnant mothers habitually compensate positionally to fetal load by extending the lower back. Our longitudinal study of 19 pregnant human females shows that adjustments to lumbar lordosis permit mothers to maintain a stable anteroposterior position of the COM as gestation progresses and fetal mass increases (Fig. 1e). Although full-term females extend their hips only slightly (about $5.6^\circ \pm 2^\circ$ (mean ± s.d.)), they extend their lower back by as much as 28° ($18^\circ \pm 10^\circ$), which realigns the COM above the hips and support base (Fig. 1e). When gravid females are experimentally constrained from exaggerating their lumbar lordosis, the COM translates by 3.2 ± 1.1 cm ($P < 0.0001$) by the end of gestation, increasing the upper body's torque around the hip roughly eightfold (Fig. 1c, d). However, when free to self-select their positional alignment, pregnant females naturally increase their lumbar lordosis, limiting anteroposterior translation of the COM within a narrow range, less than 0.3 ± 0.7 cm ($P = 0.5695$) by term (Fig. 1e). Once obstetric load has reached a threshold of about 40% of the expected term fetal mass (Fig. 2a), this lordotic adjustment increases in relation to fetal mass ($r = 0.9732$, $P = 0.0011$), thus maintaining a stable position of the COM throughout pregnancy (Fig. 2b). Extension of the lower back helps control COM position but exerts a biomechanical cost to gravid mothers in the form of shearing forces caused by the nearly 60% increase in lumbar lordosis, from a mean angle of $32^\circ \pm 12^\circ$ in early pregnancy to $50^\circ \pm 12^\circ$ at term (Fig. 2b). Two measures of the deleterious effects of spinal shearing are the increased risk of forward displacement of the lumbar vertebrae[9,10] and the higher incidence of lower back pain in pregnant women[11,12]. Greater shearing occurs because increases in lordosis transmit relatively more spinal load

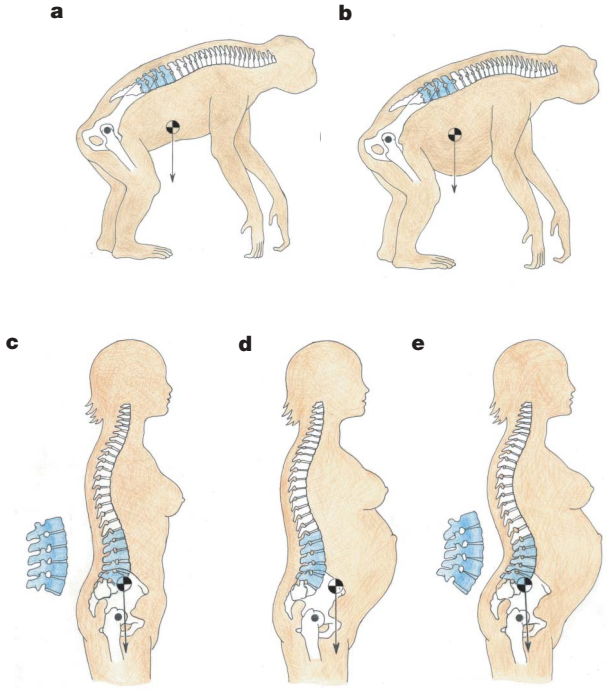


Figure 1 | COM and lumbar lordosis during pregnancy. **a**, Quadrupedal chimpanzee, non-pregnant. **b**, Quadrupedal chimpanzee, pregnant with no change in sagittal position of the COM with respect to the postural support base. **c**, Bipedal human female with typical lumbar lordosis and COM in approximate sagittal alignment with the hip. At a given 0.005-m COM distance from the hip, a 409-N upper body generates 2 N m torque at the hip. **d**, Pregnant human female with anteriorly translated COM, lacking positional adjustment of lumbar lordosis. The force of gravity, when more distant from the hip, generates a larger hip moment and an unstable upper body. With pregnancy, a 511-N upper body and a COM at 0.032 m from the hip increases the torque to 16 N m. **e**, Typical pregnant human female with naturally extended back and recovered COM by means of increased lumbar lordosis, a stable positional alignment with reduced hip torque (1.5 N m) but with exacerbated spinal shearing load. Open circle with cross hairs, COM in sagittal plane; filled circle, hip position in sagittal plane; arrow, direction of gravitational force.

[1]Department of Anthropology, Harvard University, 11 Divinity Avenue, Cambridge, Massachusetts 02138, USA. [2]Department of Anthropology, University of Texas at Austin, 1 University Station, Austin, Texas 78712, USA.

along the dorsal pillar of the spine comprising the zygapophyseal joints[13] (Fig. 3d). Typical bipedal posture directs only 16% of the total compressive load through these joints[14], and slight extension of the lower back redirects another 12% to the zygapophyses[15]. During pregnancy the mean lordotic excursion of 18° shifts even greater loads onto the zygapophyses, as much as 20–40% according to published models[15,16].

Given the demands of fetal load and the importance of pregnancy for fitness, one predicts that natural selection has operated on the unique anatomy of the hominin lumbar region to mitigate the biomechanical problems that females confront. Our analyses show that humans are characterized by a strong, derived pattern of lumbar sexual dimorphism that is evident in several aspects of the lumbar vertebrae. One major feature of human lumbar sexual dimorphism is the degree and pattern of dorsal wedging that forms the lumbar lordosis and results from a disproportionately short dorsal margin of the vertebral body. Vertebral wedging differs significantly between human sexes from L1 to L4 ($P < 0.0001$ to $P < 0.008$; Fig. 3a). The complete lordotic sequence of dorsal wedging in males spans just two vertebrae, the penultimate and last lumbar vertebrae. In contrast, the female pattern of dorsal wedging includes three vertebrae, the prepenultimate, penultimate and last lumbar vertebrae (Fig. 3a, d). This 3:2 wedging dimorphism occurs regardless of variation in the total number of lumbar elements, whether variant L4, modal L5 or variant L6, but is entirely absent in chimpanzees (Supplementary Tables 1 and 2). Females benefit from the third wedging level during pregnancy because it enables them to increase the lordosis with less intervertebral rotation. An equivalent angular excursion between L3 and L4 results in greater extension of the upper body in females than in males (Fig. 3e). In this way, females minimize shear force across lumbar vertebral joints by about 30% (Supplementary Information).

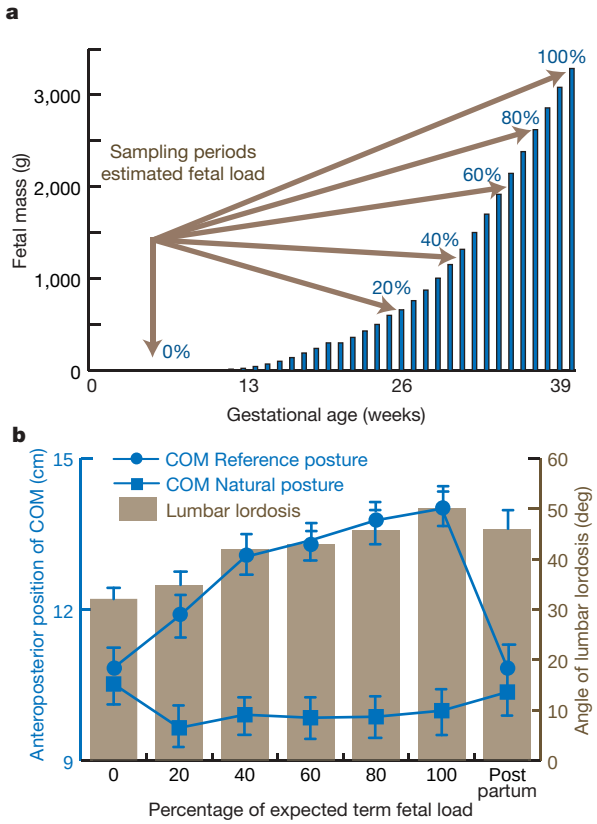


Figure 2 | Maternal COM and lumbar lordosis relative to fetal load. **a**, Increase in fetal body mass by weeks of gestation, showing six prepartum sampling sessions of sequential periods of 20% fetal load (estimated fetal mass from ref. 20). Note the increasing rate of increase in fetal mass within the second trimester and the maximum increase in the third trimester. **b**, Angle of lumbar lordosis and position of the COM with respect to human pregnancy. Means are plotted against stages of fetal growth and an approximate eight-week postpartum period. Results support the predicted relationship between COM reference posture (circles) and lordosis (bars), their strong correlation ($r = 0.9732$, $P = 0.0011$) and the constancy of COM natural posture (squares) when gravid females self-select their angle of lumbar lordosis. Circles plot the mean forward position of the COM recorded in a reference posture (see Methods) in which pregnant females were constrained from self-selecting their postural alignment. In the absence of positional adjustment, COM translates 3.2 cm from 11 cm at 0% fetal mass to 14 cm at 100% fetal mass. Note the return to the baseline position postpartum. Bars plot the mean angle of lumbar lordosis self-selected by pregnant females in natural stance. Lumbar lordosis increases from an angle of 32° at 0% fetal mass to 50° at 100% fetal mass, late in pregnancy. The angle of lumbar lordosis begins to decrease postpartum. Squares plot the resultant forward position of the COM self-selected by pregnant females in natural posture, when postural alignment was not artificially constrained. As women naturally increased their lumbar lordosis, their COM remained relatively stable, translating by no more than 1 cm during pregnancy. The difference in forward position of the COM from early pregnancy to term was 0.3 cm. $n = 19$. Data are means ± s.e.m.

Two additional key features of human lumbar sexual dimorphism are present within the dorsal pillar. First, the zygapophyseal surface area is 14% ± 3% ($P < 0.01$) larger relative to vertebral size in females than in males (Fig. 3b), which is consistent with the redirection of a larger proportion of spinal load along the dorsal structures during human pregnancy. Second, female prezygapophyseal joint surfaces are oriented more coronally by an average of 13% ± 5% ($P < 0.05$) than those of males (Fig. 3c), enhancing resistance to large shearing forces imposed by fetal mass and back extension. As in wedging, these zygapophyses are not significantly dimorphic in chimpanzees (Supplementary Table 2). In bracing the zygapophyses more coronally, human female vertebrae achieve greater buttressing against anterior displacement of vertebral bodies within the deep lumbar curve.

The evidence for lumbar sexual dimorphism in humans which improves maternal performance in posture and locomotion suggests that the distinctive hominin lumbar curve has been subject to strong selection pressures. If so, one expects these adaptations to be present in the genus *Australopithecus*, which is known to have been habitually bipedal at least two million years after the earliest bipedal hominins[2,3]. It is intriguing that, of the two nearly complete known australopith lumbar segments, Sts 14 and Stw 431, the former has the typical human female pattern with three dorsally wedged vertebrae, whereas the latter has a more male-like pattern with fewer lordotic vertebrae (Fig. 4a). One possible explanation for this difference is that one female and one male *A. africanus* are sampled. This inference is supported by the observation that the prezygapophyses of Sts 14 (L1–L6) are angled 9–12° more coronally than the measurable facets of Stw 431 (L3, L5 and L6; Fig. 4b), as is typical of the human female and male patterns, respectively (Fig. 3c). Australopiths not only had a lumbar lordosis with human-like wedging patterns, but they also had relatively large zygapophyseal facets[5] with angular dimorphism similar to that in modern humans. Because these features have a fundamental role in resisting shear force[14], similar patterns of lumbar dimorphism in *Australopithecus* and *Homo* indicate that spinal shear was also a major challenge in australopiths in general, and especially for gravid females. Similarities in body size and life history between australopiths and chimpanzees suggest that term mass and duration of gestation for australopiths was chimpanzee-like (1,590 g at 230 days)[17,18] rather than human-like (3,200 g at 290 days)[19,20]. Even so, term mass of the australopithecine fetus would easily have exceeded the 40% load trigger of 1,200 g in human pregnancy for a substantial period of pregnancy, approximately the last trimester (Supplementary Information).

Since the discovery of the first australopithecine postcrania[4] there has been a concerted study of the evolution of hominin locomotion, yet without consideration of the biomechanical challenges posed by

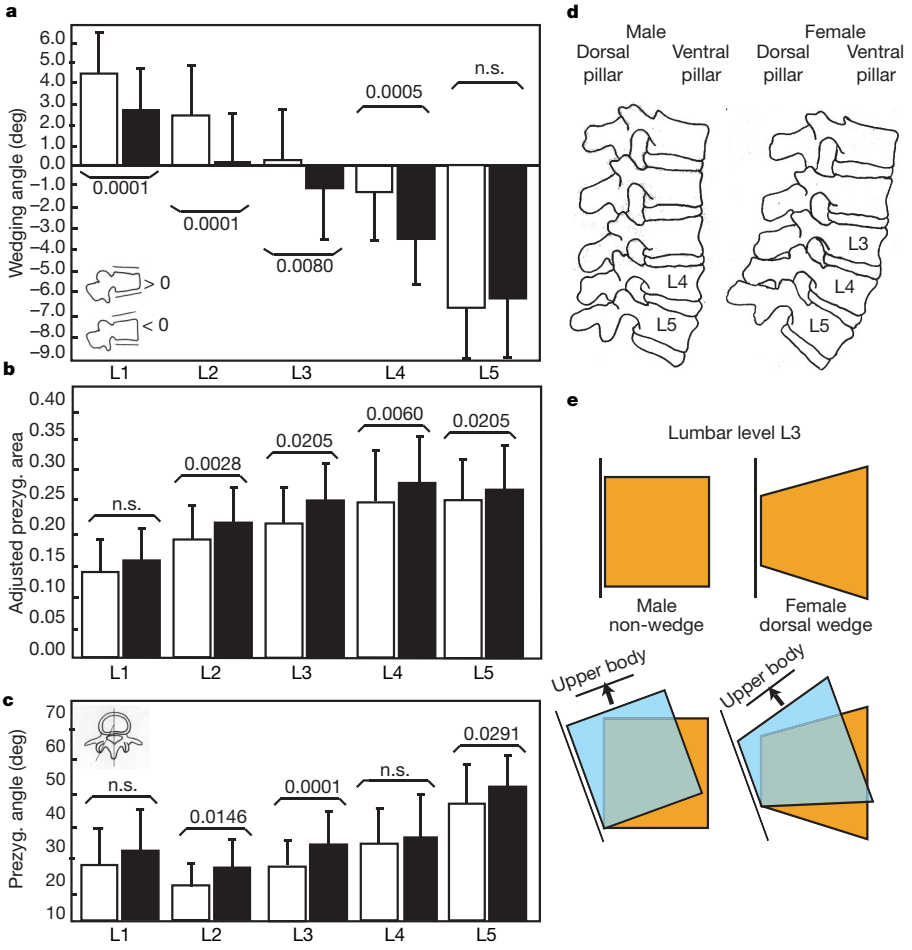


Figure 3 | Sex differences in the lumbar vertebral column of human males and females. Female values are shown by filled bars, male values by open bars. **a**, Wedging angle of vertebral bodies, angles greater than 0° are kyphotic (thoracic-type wedging), whereas angles less than 0° are lordotic (lumbar-type wedging). Females present a longer series of dorsally wedged vertebrae; L3, L4 and L5, whereas males are lordotic at only two levels, L4 and L5. **b**, Prezygapophyseal (prezyg.) area, adjusted by geometric mean for overall vertebral size. The female area is significantly larger than the male area at L2, L3, L4 and L5, indicating that females bear a greater proportion of spinal load along the dorsal pillar, which is consistent with fetal loading patterns identified during pregnancy. **c**, Prezygapophyseal angle. The female facets are significantly more oblique at L2, L3 and L5, conferring greater resistance to forward displacement of lumbar vertebrae. In **a**–**c**, $n = 59$ males, 54 females. Data are means and s.d. **d**, Diagram of lumbar region in males and females, showing contrasting mean wedging patterns and anatomical structures within the dorsal pillar (including zygapophyses) and ventral pillar (vertebral bodies). **e**, Difference in vertebral body shape in males and females at L3. There are equivalent angles of excursion yet there is greater upper body extension in the female spine. The inherent dorsal wedging shape of the female L3 relative to the non-wedged male L3 generates less shearing force when the upper body is repositioned by means of lower back extension, as occurs during fetal loading.

pregnancy. Our analyses not only show that the derived dimorphism of the lumbar lordosis in modern humans helps mothers to mitigate the shearing forces generated by fetal load, but also indicate that the biomechanical demands of pregnancy exerted an early selection pressure on the evolution of lumbar lordosis in bipedal hominins. These results highlight the vulnerability of the lumbar vertebrae to various forms of loading in bipeds, and the importance of adaptations in both the lumbar vertebrae and the dimensions of the pelvic canal[21–23] to female reproductive success. It is reasonable to hypothesize that fatigue and pain in the lower back muscle affected early hominin mothers just as they do modern mothers, possibly limiting foraging efficiency and the ability to escape from predators, leaving the gravid female at risk of nutritional stress and injury or death. Later hominins underwent a reduction in the number of lumbar vertebrae, from six to five modal vertebrae[5,24,25], along with relative increases in vertebral body size[26] possibly for carrying[5], increased trekking[27] and/or endurance running[28]. Regardless of the varied selection pressures behind these shifts, fetal load remained a persistent selection factor in the evolution of lumbar sexual dimorphism in hominins.

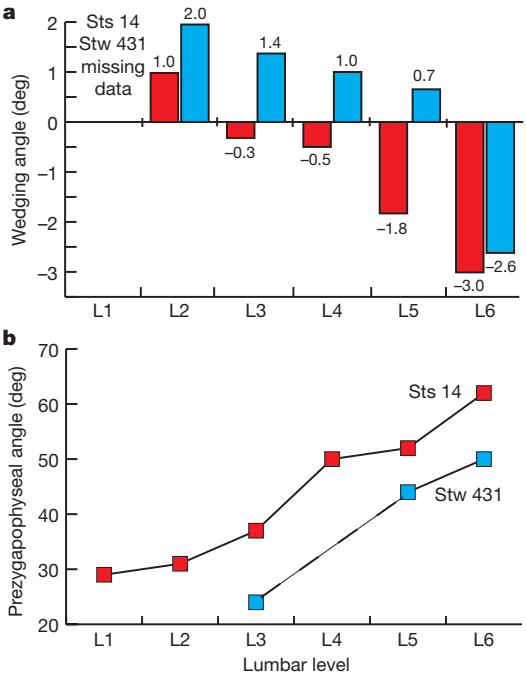


Figure 4 | Australopithecine lumbar lordosis and prezygapophyseal angle. **a**, Angle of lumbar vertebral wedging for *Australopithecus africanus* specimens Sts 14 (red) and Stw 431 (blue). Sts 14 shows a wedging pattern similar to that in modern human females, comprising the three caudalmost lumbar vertebrae, L4, L5 and L6. Although the preserved lumbar column of Stw 431 is less complete than that of Sts 14, the caudalmost levels are preserved well enough to identify a different wedging pattern. The dorsal wedging sequence of Stw 431 includes only one lumbar vertebra, at the last lumbar level. In this manner, Stw 431 is unlike Sts 14 and modern human females and is more similar to modern human males in having a shorter region of lordotic lumbar vertebrae. **b**, The prezygapophyseal angle of the preserved lumbar region for Sts 14 and Stw 431. The larger angles of Sts 14 relative to those of Stw 431 mirror the modern human female–male pattern in that Sts 14 presents more oblique angles and therefore greater coronal orientation of the prezygapophyseal facets than Stw 431 (see Supplementary Information).

METHODS SUMMARY

The anteroposterior position of the maternal COM was identified from ground reaction force vectors measured by a triaxial transducing force plate, following the zero-point-to-zero-point integration technique[29].

Angular excursions of the lumbar spine were calculated from three-dimensional positional data acquired from a Vicon motion analysis system capturing infrared reflections from surface markers that were externally adhered to palpable landmarks of the thoracic, lumbar and sacral vertebrae.

Comparative morphometrics were used to evaluate patterns of sexual dimorphism in the human lumbar spine. Linear and angular dimensions of lumbar vertebrae were measured to identify the relative size and shape of vertebral features subject to the biomechanical stresses generated by fetal load.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 21 June; accepted 3 October 2007.

1. Darwin, C. *The Descent of Man* (John Murray, London, 1871).
2. Galik, K. *et al.* External and internal morphology of the BAR 1002′00 *Orrorin tugenensis* femur. *Science* **305,** 1450–1453 (2004).
3. Zollikofer, C. P. E. *et al.* Virtual cranial reconstruction of *Sahelanthropus tchadensis*. *Nature* **434,** 755–759 (2005).
4. Robinson, J. T. *Early Hominid Posture and Locomotion* (Univ. of Chicago Press, Chicago, 1972).
5. Sanders, W. J. Comparative morphometric study of the australopithecine vertebral series Stw-H8/H41. *J. Hum. Evol.* **34,** 249–302 (1998).
6. Latimer, B. & Ward, C. V. in *The Nariokotome* Homo erectus *Skeleton* (eds Walker, A. & Leakey, R.) 266–293 (Harvard Univ. Press, Cambridge, MA, 1993).
7. Strassmann, B. I. The biology of menstruation in *Homo sapiens*: total lifetime menses, fecundity, and nonsynchrony in a natural-fertility population. *Curr. Anthropol.* **38,** 123–129 (1997).
8. Jensen, R. K., Doucet, S. & Treitz, T. Changes in segment mass and mass distribution during pregnancy. *J. Biomech.* **29,** 251–256 (1996).
9. White, A. A. & Punjabi, M. M. *Clinical Biomechanics of the Spine* (Lippincott, Philadelphia, 1990).
10. Bogduk, N. *Clinical Anatomy of the Lumbar Spine and Sacrum* (Churchill Livingstone, New York, 1997).
11. Ostgaard, H. C., Andersson, G. B. J., Schultz, A. B. & Miller, J. A. A. Influence of some biomechanical factors on low-back pain in pregnancy. *Spine* **18,** 61–65 (1993).
12. Dumas, G. A., Reid, J. G., Griffin, M. P. & McGrath, M. J. Exercise, posture, and back pain during pregnancy. Part 1. Exercise and posture. *Clin. Biomech.* **10,** 98–103 (1995).
13. Pal, G. P. & Routal, R. V. Transmission of weight through the lower thoracic and lumbar regions of the vertebral column in man. *J. Anat.* **152,** 93–105 (1987).
14. Adams, M. A. & Hutton, W. C. The effect of posture on the role of the apophyseal joints in resisting intervertebral compressive forces. *J. Bone Joint Surg. Br.* **62,** 358–362 (1980).
15. Lorenz, M., Patwardhan, A. & Vanderby, R. Jr. Load-bearing characteristics of lumbar facets in normal and surgically altered spinal segments. *Spine* **8,** 122–130 (1983).
16. Dunlop, R. B., Adams, M. A. & Hutton, W. C. Disc space narrowing and the lumbar facet joints. *J. Bone Joint Surg. Br.* **66,** 706–710 (1984).
17. Lee, D. R., Kuehl, T. J. & Eichberg, J. W. Real-time ultrasonography as a clinical and management tool to monitor pregnancy in a chimpanzee breeding colony. *Am. J. Primatol.* **24,** 289–294 (1991).
18. DeSilva, J. & Lesnik, J. Chimpanzee neonatal brain size: Implications for brain growth in *Homo erectus*. *J. Hum. Evol.* **51,** 207–212 (2006).
19. Institute of Medicine of the National Academies. *Nutrition During Pregnancy. Part 1. Weight Gain* (National Academy Press, Washington DC, 1990).
20. Alexander, G. R., Himes, J. H., Kaufman, R. B., Mor, J. & Kogan, M. A United States national reference for fetal growth. *Obstet. Gynecol.* **87,** 163–168 (1996).
21. Washburn, S. L. Sex differences in the pubic bone. *Am. J. Phys. Anthropol.* **6,** 199–207 (1948).
22. Schultz, A. H. Sex differences in the pelves of primates. *Am. J. Phys. Anthropol.* **7,** 401–423 (1949).
23. Rosenberg, K. & Trevathan, W. Bipedalism and human birth: The obstetrical dilemma revisited. *Evol. Anthropol.* **4,** 161–168 (1996).
24. Arensburg, B. in *Le Squelette Mousterien de Kebara 2* (eds Bar Yosef, O. & Vandermeersch, B.) 113–146 (Cahiers de Paléoanthropologie, Paris, 1991).
25. Trinkaus, E. *The Shanidar Neandertals* (Academic, New York, 1983).
26. Jungers, W. L. Relative joint size and hominoid locomotor adaptations with implications for the evolution of hominid bipedalism. *J. Hum. Evol.* **17,** 247–265 (1988).
27. Ruff, C. B. in *Primate Locomotion: Recent Advances* (eds Strasser, E., Fleagle, J., Rosenberger, A. & McHenry, M.) 449–469 (Plenum, New York, 1998).
28. Bramble, D. M. & Lieberman, D. E. Endurance running and the evolution of *Homo*. *Nature* **432,** 345–352 (2004).
29. Zatsiorsky, V. & King, D. An algorithm for determining gravity line location from posturographic recordings. *J. Biomech.* **31,** 161–164 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank W. Sanders for fossil measurements and discussion; S. Ford, J. Jensen, J. Kappelman, D. Overdorff, D. Pilbeam, D. Raichlen, P. Rightmire and C. Ruff for comments and assistance with research; L. Gordon, D. Hunt, L. Jellema, B. Latimer and R. Thorington for access to specimens; and the Developmental Motor Control Laboratory at the University of Texas, Austin, for laboratory use. Figure 1 was drawn by L. Meszoly. This work was supported by grants from the National Science Foundation (to L.J.S. and K.K.W.), the L. S. B. Leakey Foundation (to K.K.W.), the National Science Foundation (to D.E.L.) and the American School of Prehistoric Research (Harvard).

Author Contributions K.K.W. designed the study, conducted the experiments, and analysed and interpreted the data. L.J.S. assisted in the study design and the interpretation of results. D.E.L. assisted in the fossil study design and in the analysis and interpretation of the fossil and biomechanical data. K.K.W., L.J.S. and D.E.L. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to K.K.W. (whitcome@fas.harvard.edu).

METHODS

Kinematic/kinetic sample structure. Nineteen pregnant women between the ages of 20 and 40 years participated in the longitudinal study, initiated at the third month of pregnancy and concluded in the third month of post parity. Study protocol received University of Texas at Austin IRB approval for human research. Volunteers were excluded if they demonstrated life histories characterized by joint illness/injury or previous pregnancy-related difficulties leading to medical treatment, restricted physical activity, or persistent discomfort. Maternal body weight was recorded each session and assessed by the Institute of Medicine standards[19], which recommend an increase of 1.36–1.81 kg in the first three months and 1.36–1.81 kg per month in the later trimesters. Subjects whose prenatal weight gain exceeded 12.75 kg would have been excluded from the analyses, but none eclipsed the parameter. To ensure that comparisons across subjects matched successive stages of fetal load, data collection sessions targeted seven parity windows of 0%, 20%, 40%, 60%, 80% and 100% fetal mass and a final session postpartum.

All kinematic and kinetic data were collected in the Developmental Motor Control Laboratory at the University of Texas at Austin. A Vicon motion analysis system (Vicon Peak) captured three-dimensional positional data (60 Hz sampling rate) of each subject during quiet stance and while walking freely though a 2 m^3 viewing volume. Five infrared cameras recorded positional data and trajectories of lightweight 25-mm reflective markers externally adhered over spinous processes of vertebrae L1 (lumbar level 1), L4 (lumbar level 2) and S2 (sacral level 2), identified by palpation. The time reference of heel strike and toe-off was identified by the onset and cessation of vertical force, respectively, as registered on a triaxial transducing force plate (600 Hz sampling rate). Before each data collection session the viewing volume was calibrated by following static and dynamic protocols. Residuals for all cameras were consistently within a range of 0.400 to 0.594 mm, representing less than 0.1% of the 2 m^3 viewing volume. The mean wand visibility approached 84.0%.

Kinematic/kinetic measurements. Vicon three-dimensional data files were transferred to a personal computer on which the lordotic angle was calculated algorithmically from positional data derived from lumbar vertebrae with BodyBuilder software (Vicon Peak). Angles were exported as ASCII to Microsoft Excel files for further analysis.

Three points defined by the vertebral markers L1, L4 and S2 allowed quantification of the lordotic angle between segments 1 and 2 defined by markers L1–L4 and L4–S2, respectively. Larger angles indicated more acute lumbar lordosis.

Kinematic and force-plate analogue data were captured to calculate the maternal total body COM in both the reference and self-selected postures. The static measure of COM taken in the consistent reference posture was needed to identify the translation of the resultant COM. Angular changes in lumbar lordosis were assessed functionally relative to the translation of this reference posture COM. To obtain as consistent a reference posture COM as possible, a portable plywood wall 3 feet × 6 feet (about 91 cm × 183 cm) was supported above the floor on a wheeled assembly spanning the force plate. Subjects stood with head, shoulders and buttocks in contact with the vertical panel. Once a stable posture had been attained, the portable wall was retracted. A second static measure of maternal COM was taken during natural stance to determine any self-selected kinematic repositioning of the COM. Reference posture COM was predicted to change significantly during pregnancy, as the segmental angles of lumbar lordosis and pelvic tilt were held constant from session to session through postural alignment with the reference panel. In contrast, the self-selected position of the maternal COM was expected to remain relatively constant throughout the study, its stability achieved through natural adjustments in lumbar lordosis.

The fore–aft vectors of the ground reaction force and centre of pressure from which COM values were calculated were recorded with a Bertec K70501 type 4550-08 force plate located in the centre of an open laboratory space, allowing subjects to achieve natural postures. Maternal body mass was recorded from the force plate as the z force component adjusted for the plate's baseline measure taken during the corresponding session.

To obtain the maternal COM during both reference posture and natural stance, the horizontal position of the static centre of gravity was calculated from vectors measured by the force plate by using the zero-point-to-zero-point integration technique introduced by Zatsiorsky & King[29], with the formula

$$X_{\mathrm{GLP}}(t) = \left[\int\int_{t_n|-\delta<F_x<\delta}^{t_n+1|-\delta<F_x<\delta} \ddot{X}(t) \right] + \dot{X}(t_n)t + X_{\mathrm{COP}}(t_n)$$

where $X_{\mathrm{GLP}}(t)$ is the horizontal position of the static centre of gravity, t_n is time n, the vertical bar stands for 'under the condition that', F_x is the horizontal ground reaction force, δ is the incremental value, $\ddot{X}$ is acceleration, $\dot{X}$ is velocity and X_{COP} is the centre of pressure location along the x axis.

The method is based on the postulation that the horizontal position of the total body line of gravity and the total body centre of pressure on the force plate coincide when the horizontal ground reaction force, F_x, is zero. At this instant the torque about the intersection between the vertical axis through the ankles and the supporting substrate is either zero or negligible. The algorithm used to calculate the position of the COM was validated by Zatsiorsky and King[29] with videography-based segment mass. There was no significant difference (at the 0.05 level) and coefficients of correlation were high (0.79–0.96) (ref. 29).

The position of the maternal COM in both the reference posture and the natural stance was determined relative to a point of reference. The reference posture served to target a rigid anatomical reading of the position of the COM. The C7 marker was expected to be the most relevant and accurate body marker for calculation of the reference posture COM position, because it is the marker least likely to shift directionally in anatomical position relative to the location of the COM (among the non-dependent variable markers). Because the torso is a relatively solid segment, the C7 marker, adhered to the external palpable spinous process of the seventh cervical vertebra, provided a consistent reference for determining the fore–aft position of the maternal COM in the experimental condition on the reference board. In order to calculate the position of the COM during natural stance, the heel marker representing the base of support was used as a point of reference.

Repeated-measures analysis of variance (ANOVA, time × condition) was used to assess whether maternal gait kinematics and maternal COM differed with incremental increases in fetal growth. Both linear and nonlinear models were included because mass increase during pregnancy is nonlinear[8]. Repeated-measures design is appropriate for longitudinal data of this type, by providing a more precise estimate of the experimental error. The technique identifies variability due to individual differences because the same subjects take part in each condition. Because the variance caused by differences between individuals is not helpful in deciding whether there is difference between occasions, the known individual differences can be isolated from the analysis by subtraction from the error variance. This step increases the power of the analysis. Repeated-measures ANOVA models correlation between the repeated measures, which is important because the longitudinal series violates assumptions of independence. To test for the presence of significant differences in dependent variables at early-stage fetal load and at term fetal load at the group level, the non-parametric Wilcoxon rank sums test was applied. Statistical significance for the analyses was determined a priori at a level of $P \leq 0.05$ for the independent variable of fetal load and three dependent variables of maternal COM and maternal lumbar lordosis angle. Adjustments for repeated tests were made with the Bonferroni correction.

Morphometric sample structure. The sample population of 59 males and 54 females chosen to test the study hypothesis was drawn from two well-studied twentieth-century osteological archives of known age and sex: the Hamann–Todd collection, curated at the Cleveland Museum of Natural History, and the Terry collection, housed at the National Museum of Natural History in Washington DC. Ancestry-related differences within the sample population (morgue identified) were tested for ethnicity effect by using ANOVA cross (sex and ethnicity). No significant ethnicity response by sex was obtained.

Autopsy records and morgue photos were examined to identify sex. All specimens were further assessed for sex in accordance with the modified Phenice method[30,31]. Individuals whose sex was ambiguous according to either collection records or observer Phenice assessment were excluded.

Specimens were selected within an adult age range of 20–40 years. This criterion targeted individuals whose skeletal development had reached maturity but whose ageing effects had not yet eclipsed osteophytic deposition, typical in synovial and symphyseal joint margins with ageing, for example spondylosis deformans[10]. Chronological age was obtained through morgue records and further evaluated by visual confirmation of postcranial epiphyseal fusion. If skeletal age was found to fall outside the inclusion range, the specimen was omitted from the study. Pathological specimens, whether determined by collection records or gross observation, were not analysed.

Lumbar vertebrae were defined in accordance with their zygapophyseal orientation[32,33]. This facet-based designation differs from the widely used non-rib-bearing alternative[34] in its functional emphasis on the range of motion between vertebral elements; type and range of movement in the lumbar column are largely influenced by facet direction. The medial and lateral orientation of lumbar superior and inferior facets, respectively, guide sagittal flexion and extension while resisting both rotation[35] and ventral displacement[6,10].

Lumbar osteological measurements and analyses. Predictions of lumbar vertebral sexual dimorphism were tested on 14 vertebral variables at each lumbar vertebral level, chosen to define the relative size and shape of the lumbar vertebrae. Linear measurements were collected with a Mitutoyo 500-171 needle-point digital calliper and were recorded to the nearest 0.01 mm. Angular measurements were collected with an SPI 0–180° protractor.

Prezygapophyseal surface area was calculated from geometric mean adjusted linear variables using the equation for an ellipse. Linear measurements were used to calculate an angular variable of vertebral body wedging as described by Digiovanni *et al.*[36]:

$$\text{Wedging angle} = 2\arctan\{[(\text{centrum dorsal height} - \text{centrum ventral height})/2]/\text{centrum anteroposterior diameter}\}$$

Positive angles were kyphotic; negative angles were lordotic. A vertebra was determined to be neutral—neither kyphotic nor lordotic—when its value fell within the range 0.5° to -0.5°. JMP 5.0.1.2 (SAS Institute) and SPSS 12.0 (SPSS, Inc.) software packages were used for statistical analyses.

Without adjustment for body size variation within the sample population, any significant differences identified by contrasting males and females might reflect little more than stochastic distribution of body size differences within the samples. The representative measure of gross size used to remove the general isometric phenomenon[37,38] was the scale-free geometric mean[39,40] derived from the 48 linear variables of the lumbar vertebrae, 12 from each of the first, second, penultimate and last lumbar levels. Mosimann's[39] method removes the effects of size for each variable on an individual basis, using a directly measured index of individual size. Variates obtained for each individual were standardized by dividing the raw values by the geometric mean of the relevant specimen (the 48th root of the product of the variables).

In accordance with the biomechanical principles outlined in the two-pillar model of spinal force transmission[13] the variables represent the major load-bearing and load-resistant structures operating under conditions of bipedal obstetric load. Variables were tested for normality with the single-sample Shapiro–Wilk *W* test. A between-sex test for homoscedasticity was performed as a two-tailed F_{max} test with a 0.05 α. Because distribution assumptions of normality and homoscedasticity were not met for many of the variates, tests of significance in comparing male and female specimens were obtained with the Wilcoxon rank sums test using a multiple-comparisons adjustment to limit type I errors[41] as described by Jaccard and Wan[42], who advocated a modified Bonferroni procedure. The Wilcoxon rank sums test is a non-parametric test of the null hypothesis that both male and female samples for each variable derive from the same distribution.

30. Phenice, T. W. A newly developed visual method of sexing in the *Os pubis*. *Am. J. Phys. Anthropol.* **30,** 297–301 (1969).
31. Ubelaker, D. H. & Volk, C. G. A test of the Phenice method for the estimation of sex. *J. Forensic Sci.* **47,** 19–24 (2002).
32. Washburn, S. L. & Buettner-Janusch, J. The definition of thoracic and lumbar vertebrae. *Am. J. Phys. Anthropol.* **10,** 251–252 (1952).
33. Shapiro, L. in *Postcranial Adaptation in Nonhuman Primates.* (ed. Gebo, D.L.) 121-149 (Northern Illinois University Press, DeKalb, IL, 1993).
34. Schultz, A. H. The skeleton of the trunk and limbs of higher primates. *Hum. Biol.* **2,** 303–438 (1930).
35. Rockwell, H., Gaynor Evans, F. & Pheasant, H. The comparative morphology of the vertebrate spinal column: its form as related to function. *J. Morphol.* **63,** 87–117 (1938).
36. Digiovanni, B., Scoles, P. & Latimer, B. Anterior extension of the thoracic vertebral bodies in Scheuermann's kyphosis: an anatomic study. *Spine* **14,** 712–716 (1989).
37. Corruccini, R. S. Shape in morphometrics: comparative analyses. *Am. J. Phys. Anthropol.* **73,** 289–303 (1987).
38. Jungers, W. L., Falsetti, A. B. & Wall, C. E. Shape, relative size, and size-adjustments in morphometrics. *Yb. Phys. Anthropol.* **38,** 137–161 (1995).
39. Mosimann, J. Size allometry: Size and shape variables with characterizations of the log normal and gamma distributions. *J. Am. Stat. Assoc.* **65,** 930–945 (1970).
40. Darroch, J. & Mosimman, J. Canonical and principal components of shape. *Biometrika* **72,** 241–252 (1985).
41. Sokal, R. R. & Rohlf, F. J. *Biometry: The Principles and Practice of Statistics in Biological Research.* 3rd Edition. (W.H. Freeman and Company, New York, 1995).
42. Jaccard, J. & Wan, C. K. *LISREL Approaches to Interaction Effects in Multiple Regression* (Sage, Thousand Oaks, CA, 1996).

LETTERS

Coevolution with viruses drives the evolution of bacterial mutation rates

Csaba Pal[1,2], María D. Maciá[3], Antonio Oliver[3], Ira Schachar[1] & Angus Buckling[1]

Bacteria with greatly elevated mutation rates (mutators) are frequently found in natural[1–3] and laboratory[4,5] populations, and are often associated with clinical infections[6,7]. Although mutators may increase adaptability to novel environmental conditions, they are also prone to the accumulation of deleterious mutations. The long-term maintenance of high bacterial mutation rates is therefore likely to be driven by rapidly changing selection pressures[8–14], in addition to the possible slow transition rate by point mutation from mutators to non-mutators[15]. One of the most likely causes of rapidly changing selection pressures is antagonistic coevolution with parasites[16,17]. Here we show whether coevolution with viral parasites could drive the evolution of bacterial mutation rates in laboratory populations of the bacterium *Pseudomonas fluorescens*[18]. After fewer than 200 bacterial generations, 25% of the populations coevolving with phages had evolved 10- to 100-fold increases in mutation rates owing to mutations in mismatch-repair genes; no populations evolving in the absence of phages showed any significant change in mutation rate. Furthermore, mutator populations had a higher probability of driving their phage populations extinct, strongly suggesting that mutators have an advantage against phages in the coevolutionary arms race. Given their ubiquity, bacteriophages may play an important role in the evolution of bacterial mutation rates.

Antagonistic coevolution with parasites (the reciprocal evolution of host defence and parasite counter-defence mechanisms) has long been recognized as a potentially important force in the evolutionary maintenance of sexual reproduction[16]. Sex often results in offspring that are genetically distinct from their parents, and hence may be more resistant to parasites that are adapted to parental genotypes[16]. Increased mutation rates in bacterial populations may confer similar indirect benefits in the absence of sex. Lytic bacteriophages are ubiquitous and require bacterial cell lysis after infection and replication to transmit to new hosts; hence there is very strong reciprocal selection for bacterial resistance and phage infectivity. This interaction can lead to ongoing antagonistic coevolution between bacteria and phages[18,19], which creates conditions where mutator alleles may increase in frequency by hitch-hiking with the beneficial resistance mutations they generate.

We performed simple computer simulations to address the conditions under which coevolution with bacteriophages could result in the evolution of elevated mutation rates in bacteria, and the consequences of mutators to the fitness of the coevolving phage populations (see Supplementary Information). Despite an inherent cost to being a mutator (an increased chance of accumulating deleterious mutations at loci under stabilizing selection), mutators were able to increase in frequency under a wide range of conditions. An example of the dynamics of a mutator allele and one particular host genotype is shown in Fig. 1a. Furthermore, increasing the mutation supply rate of the bacterial population (the product of population size and mutation rate) caused a decrease in bacteriophage fitness (Fig. 1b) as a result of an increased resistance of bacteria to their contemporary phage population. The success of mutators decreased with increased costs associated with resistance to phages, increased temporal fluctuations in population sizes of bacteria and phages, and when the specificity of interaction between bacteria and phages[20,21] allowed generalists with wide resistance and infectivity ranges, respectively, to evolve.

To experimentally address whether coevolution with bacteriophages drives the evolution of mutation rates, we evolved 36 populations of the common plant-colonizing bacterium *P. fluorescens*[22] in laboratory microcosms in the presence of a naturally associated lytic DNA phage[18], and 36 populations in the absence of phages. Cultures were propagated by batch culture in King's Media B (KB), diluting 1% of each culture in fresh media on a daily basis, for a total of 24

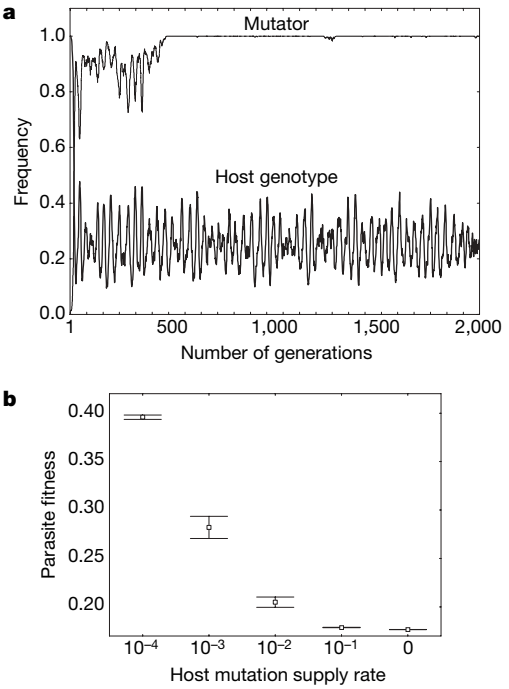


Figure 1 | Simulation results (see Supplementary Information). a, Change in frequency of the mutator through time under a Matching Alleles model ($a = 1$). Graph also shows the dynamics of one (out of four possible) host-resistance genotypes, which fluctuates through time as a result of coevolution with phages. For details on parameter set used, see Supplementary Information. **b**, Mean ± two standard errors of the mean fitness (measured over the final 500 generations of a 2,000 generation simulation) of phage populations as a function of the mutation supply rate (product of mutation rate and population size) of the host population.

[1]Department of Zoology, University of Oxford, Oxford OX1 3PS, UK. [2]Institute of Biochemistry, Biological Research Center, Temesvári krt. 62. Szeged, H-6701, Hungary. [3]Servicio de Microbiologia and Unidad de Investigación, Hospital Son Dureta, Instituto Universitario de Investigación en Ciencias de la Salud (IUNICS), 07014, Palma de Mallorca, Spain.

transfers (approximately 170 bacterial generations), and frozen every six transfers. Previous studies have shown that SBW25 and SBW25ϕ2 undergo antagonistic coevolution in KB[18,20,23], and we confirmed this result in the present study (see Supplementary Information).

We estimated the mutation rates of each population using fluctuation tests[4,24]. After 24 transfers, we found that mutation rates had increased 10- to 100-fold in 9 out of 36 populations coevolving with phages (Fig. 2). No significant increases in mutation rates were observed in any of the populations evolving in the absence of phages. We subsequently estimated mutation rates in all populations coevolving with phages at transfers 6, 12 and 18, and found a steady increase in the frequency of populations with elevated mutation rates through time (Fig. 3). We continued evolving the nine mutator populations identified at transfer 24 for a further 24 transfers and found that mutation rates remained at significantly higher levels than the ancestor in all cases.

Bacterial mutation rates can increase through phenotypic stress responses[25] as well as through genetic mutation. To confirm a genetic basis to the elevated mutation rates in our populations, we sought to identify the genes in which the mutations occurred. We isolated four clones from each of the nine mutator populations from transfer 24; of these, mutator clones were identified in seven populations (one out of four clones were mutators in three populations; three out of four in one population; and four out of four in three populations). In the other two populations, mutators must have been at relatively low frequencies. We chose a single random mutator clone from each of these seven populations for further analysis. Previous studies suggest that most mutators result from mutations in the methyl-directed mismatch repair (MMR) system[1,2,4–6], so we attempted to systematically complement MMR alleles with wild-type alleles from the closely related bacteria *P. aeruginosa*[26,27] (see Supplementary Information). In six out seven cases, wild-type mutation rates were restored by this complementation process: five with the *mutL* wild-type *P. aeruginosa* allele, and one with the *mutS* allele (see Supplementary Table 3). It is unclear which mutations were responsible for the elevated mutation rate in the seventh clone.

We have shown that mutators are more likely to evolve when bacteria coevolve with phages than when they evolve in isolation, but it is less clear why. In our simulations, mutators are indirectly favoured because they hitch-hike with alleles that confer resistance to coevolving phage populations. An increase in the frequency of mutators should therefore reduce phage fitness. Strong support for this hypothesis is that phage populations showed a greater tendency to be driven extinct when associated with mutator bacteria, compared with extinction rates of phages with non-mutators (Fig. 3; randomization test: $P = 0.015$; see Supplementary Information). Coevolution with phages could also reduce bacterial population density, which could result in the evolution of mutation rates through both an increased probability of genetic drift and reduced mutation supply rate (the product of mutation rate and population size)[28]. To address this possibility, we measured population densities of bacteria coevolving with phages across time points (transfer 6, 12, 18). We found no differences between populations with or without elevated mutation rates, strongly suggesting that population size was not an important factor contributing to patterns of mutation rate evolution (Mann–Whitney *U*-test, $P > 0.4$ for all three time points).

To confirm the benefits of elevated mutation rates of bacteria when coevolving with phages, we constructed a *mutS* knockout of *P. fluorescens* SBW25 (see Supplementary Information), which conferred an approximately 100-fold higher mutation rate. When the wild-type and *mutS* mutant were competed (see Supplementary Information), we generally found a massive selective advantage in the presence, but not the absence, of phages (Fig. 4). However, this advantage was positively frequency dependent, such that the mutators could always invade when initiated at frequencies of 10^{-2} and 10^{-4}, but that invasion success was limited when initiated at frequencies of 10^{-6}. This last result is consistent with the results of our first experiment, where mutators only increased in frequency in a quarter of populations over the course of 170 generations. Such positive frequency-dependent selection of mutators is consistent with previous theoretical[10–12] and experimental studies[5,29]. (see also simulations in the Supplementary Information). It presumably results from the wild-type population having a higher probability of evolving beneficial mutations when the mutator population is at a very low frequency. In a separate experiment, we found that populations of the *mutS* knockout were much more likely to drive their coevolving populations of phages extinct than were populations of the wild type (phage were at undetectable levels in 3 out of 24 versus 10 out of 24 replicate populations in the presence of the mutator and wild type, respectively; Fisher's exact test, $P = 0.02$). Thus these experiments initiated with mutator and wild-type genotypes confirm results from our *de novo* evolution experiments (1) that mutators are likely to have a selective advantage when coevolving with phages, and (2) that the mutators provide an advantage relative to phage in the coevolutionary arms race, presumably because of the more rapid generation of resistance mutations.

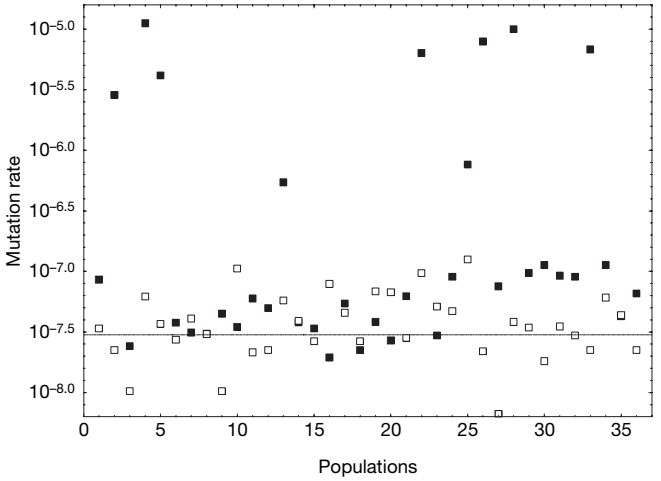


Figure 2 | Relative mutation rates. The estimated mutation rate (for rifampicin resistance) of bacteria in populations evolving with (closed symbols) and without (open symbols) phages; the line indicates the ancestor. Relative estimates using streptomycin gave the same qualitative results. Nine out of 36 populations coevolving with phages had evolved significantly higher mutation rates than the ancestor (Mann–Whitney *U*-test, $n = 6$ for evolved population and ancestral populations, $P < 0.01$ for all cases), whereas no control populations were mutators ($P > 0.1$ for all cases). The number of mutator populations was higher in the presence versus the absence of phages (Fisher's exact test: $P = 0.001$).

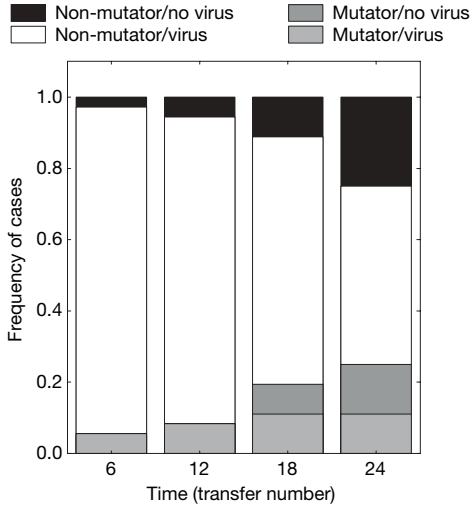


Figure 3 | The frequency of coevolving populations of bacteria evolving elevated mutation rates and driving phages extinct, through time. By transfer 24, 6 out of 9 mutator populations had driven their phages extinct, while 9 out of 27 non-mutator populations had driven their phages extinct.

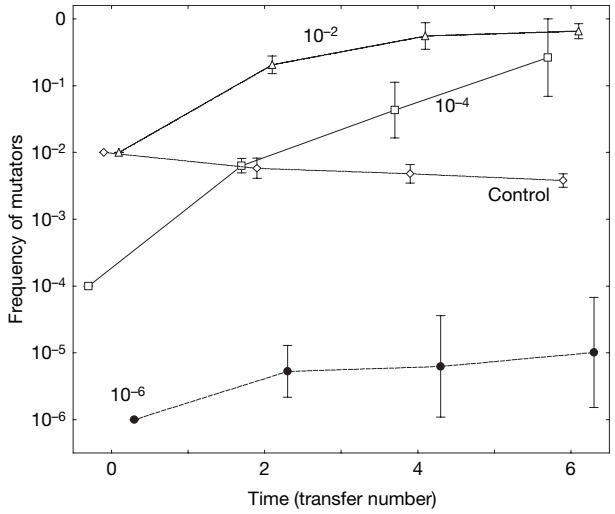


Figure 4 | Competition experiments between wild-type and isogenic mutator. Mean ± 95% confidence intervals through time. There was a significant increase in the frequency of mutators through time when initiated at frequencies of 0.01 (triangles; $F_{3,20} = 113.9$, $P < 0.0001$) and 10^{-4} (squares; $F_{3,20} = 46.1$, $P < 0.0001$), but not at 10^{-6} (filled circles; $F_{3,20} = 1.5$, $P > 0.2$ in the presence of phages. There was a significant decrease in mutator frequency in the absence of phages when mutators were initiated at 0.01 (open diamonds; $F_{3,20} = 7.1$, $P < 0.001$).

Here we have shown that antagonistic coevolution with phages can drive the evolution of elevated mutation rates in bacterial populations. The most probable explanation for this result is that mutator alleles hitch-hike with the beneficial phage resistance mutation they generated. The ubiquity of bacteriophages suggests that they may play a pivotal role in explaining why mutators persist at relatively high frequencies in many natural bacterial populations. As such, targeting phage populations may weaken selection for mutator bacteria in clinical infections. More generally, the study provides the first direct experimental evidence that a mechanism that increases genetic variation can be individually advantageous when coevolving with parasites.

METHODS SUMMARY

Study organisms and culture conditions. We coevolved the common plant-colonizing bacterium, *Pseudomonas fluorescens* SBW25 (ref. 22) and a naturally associated DNA phage, SBW25Φ2 (ref. 18). Note that we do not yet know if bacterial mutation rates affect phage mutation rates, either physiologically or by imposing selection. Seventy-two microcosms (25 ml glass universal bottle microcosms containing 6 ml King's medium B (KB)) were inoculated with 10^8 cells of *P. fluorescens* SBW25 (ref. 22). Half of the 72 microcosms were inoculated with 10^5 clonal particles of DNA phage SBW25Φ2 (ref. 18). Under these conditions, phages fail to completely lyse bacterial populations, because of the rapid emergence of bacteria resistant to the ancestral phage. Populations were propagated in a shaken incubator (200 r.p.m.; 0.9*g*) at 28 °C. Sixty microlitres of each culture was transferred to fresh medium every 24 h, for 24 transfers (approximately 170 generations). After every sixth transfer, populations were frozen in 1:4 v:v glycerol:KB solution at −86 °C.

Measurement of mutation frequency. We used modified fluctuation tests to estimate bacterial mutation rates[4,24]. Six microcosms per population were inoculated with 100–1000 bacterial cells and were allowed to grow for 24 h in a shaken (0.9*g*) 28 °C incubator. We regularly checked for the presence of pre-existing mutants of the trait investigated (antibiotic resistance) in the starting populations. Final cell density was determined by plating dilutions on non-selective solid medium (KB). The number of mutants was estimated by plating 60 μl of each culture on solid selective medium (KB plates mixed with rifampicin (100 mg ml^{-1}) or streptomycin (50 mg ml^{-1})). Jones median estimator was used to calculate mutation rate from the average and median frequency of mutant colonies[30]. Importantly, presence of phage in bacterial cells had no direct effect on mutation rates: increased mutation rates remained after isolating bacterial populations from phage using Virkon[20]. Note that Virkon treatment caused up to 40% bacterial mortality.

Received 9 July; accepted 4 October 2007.
Published online 2 December 2007.

1. LeClerc, J. E., Li, B. G., Payne, W. L. & Cebula, T. A. High mutation frequencies among *Escherichia coli* and *Salmonella* pathogens. *Science* **274,** 1208–1211 (1996).
2. Matic, I. *et al.* Highly variable mutation rates in commensal and pathogenic *Escherichia coli*. *Science* **277,** 1833–1834 (1997).
3. Trong, H. N. G., Prunier, A. L. & Leclercq, R. Hypermutable and fluoroquinolone-resistant clinical isolates of *Staphylococcus aureus*. *Antimicrob. Agents Chemother.* **49,** 2098–2101 (2005).
4. Sniegowski, P. D., Gerrish, P. J. & Lenski, R. E. Evolution of high mutation rates in experimental populations of *E. coli*. *Nature* **387,** 703–705 (1997).
5. Giraud, A. *et al.* Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut. *Science* **291,** 2606–2608 (2001).
6. Oliver, A., Canton, R., Campo, P., Baquero, F. & Blazquez, J. High frequency of hypermutable *Pseudomonas aeruginosa* in cystic fibrosis lung infection. *Science* **288,** 1251–1253 (2000).
7. Denamur, E. *et al.* High frequency of mutator strains among human uropathogenic *Escherichia coli* isolates. *J. Bacteriol.* **184,** 605–609 (2002).
8. Leigh, E. G. Natural selection and mutability. *Am. Nat.* **104,** 301–305 (1970).
9. Ishii, K., Matsuda, H., Iwasa, Y. & Sasaki, A. Evolutionarily stable mutation-rate in a periodically changing environment. *Genetics* **121,** 163–174 (1989).
10. Taddei, F. *et al.* Role of mutator alleles in adaptive evolution. *Nature* **387,** 700–702 (1997).
11. Tenaillon, O., Toupance, B., Le Nagard, H., Taddei, F. & Godelle, B. Mutators, population size, adaptive landscape and the adaptation of asexual populations of bacteria. *Genetics* **152,** 485–493 (1999).
12. Tanaka, M. M., Bergstrom, C. T. & Levin, B. R. The evolution of mutator genes in bacterial populations: the roles of environmental change and timing. *Genetics* **164,** 843–854 (2003).
13. Palmer, M. E. & Lipsitch, M. The influence of hitchhiking and deleterious mutation upon asexual mutation rates. *Genetics* **173,** 461–472 (2006).
14. Andre, J. B. & Godelle, B. The evolution of mutation rate in finite asexual populations. *Genetics* **172,** 611–626 (2006).
15. Denamur, E. *et al.* Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. *Cell* **103,** 711–721 (2000).
16. Hamilton, W. D., Axelrod, R. & Tanese, R. Sexual reproduction as an adaptation to resist parasites (a review). *Proc. Natl Acad. Sci. USA* **87,** 3566–3573 (1990).
17. West, S. A., Lively, C. M. & Read, A. F. A pluralist approach to sex and recombination. *J. Evol. Biol.* **12,** 1003–1012 (1999).
18. Buckling, A. & Rainey, P. B. Antagonistic coevolution between a bacterium and a bacteriophage. *Proc. R. Soc. Lond. B* **269,** 931–936 (2002).
19. Mizoguchi, K. *et al.* Coevolution of bacteriophage PP01 and *Escherichia coli* O157: H7 in continuous culture. *Appl. Environ. Microbiol.* **69,** 170–176 (2003).
20. Morgan, A. D., Gandon, S. & Buckling, A. The effect of migration on local adaptation in a coevolving host–parasite system. *Nature* **437,** 253–256 (2005).
21. Agrawal, A. & Lively, C. M. Infection genetics: gene-for-gene versus matching-alleles models and all points in between. *Evol. Ecol. Res.* **4,** 79–90 (2002).
22. Rainey, P. B. & Bailey, M. J. Physical and genetic map of the *Pseudomonas fluorescens* SBW25 chromosome. *Mol. Microbiol.* **19,** 521–533 (1996).
23. Morgan, A. D. & Buckling, A. Relative number of generations of hosts and parasites does not influence parasite local adaptation in coevolving populations of bacteria and phages. *J. Evol. Biol.* **19,** 1956–1963 (2006).
24. Luria, S. & Delbruck, M. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28,** 491–511 (1943).
25. Bjedov, I. *et al.* Stress-induced mutagenesis in bacteria. *Science* **300,** 1404–1409 (2003).
26. Oliver, A., Levin, B. R., Juan, C., Baquero, F. & Blazquez, J. Hypermutation and the preexistence of antibiotic-resistant *Pseudomonas aeruginosa* mutants: implications for susceptibility testing and treatment of chronic infections. *Antimicrob. Agents Chemother.* **48,** 4226–4233 (2004).
27. Oliver, A., Baquero, F. & Blazquez, J. The mismatch repair system (mutS, mutL and uvrD genes) in *Pseudomonas aeruginosa*: molecular characterization of naturally occurring mutants. *Mol. Microbiol.* **43,** 1641–1650 (2002).
28. de Visser, J., Zeyl, C. W., Gerrish, P. J., Blanchard, J. L. & Lenski, R. E. Diminishing returns from mutation supply rate in asexual populations. *Science* **283,** 404–406 (1999).
29. Chao, L. & Cox, E. C. Competition between high and low mutating strains of *Escherichia coli*. *Evolution Int. J. Org. Evolution* **37,** 125–134 (1983).
30. Rosche, W. A. & Foster, P. L. Determining mutation rates in bacterial populations. *Methods* **20,** 4–17 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Spiers for providing sequence data. This work was funded by NERC UK (A.B. and C.P.); the Royal Society (A.B.); EMBO and Hungarian Research Grant (C.P.); Ministerio de Sanidad y Consumo, Instituto de Salud Carlos III, Spanish Network for the Research in Infectious Diseases (A.O. and M.D.M.).

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.B. (angus.buckling@zoo.ox.ac.uk) or C.P. (cpal@ramet.elte.hu).

LETTERS

Initiation of zebrafish haematopoiesis by the TATA-box-binding protein-related factor Trf3

Daniel O. Hart[1,2], Tamal Raha[1,2], Nathan D. Lawson[2] & Michael R. Green[1,2]

TATA-box-binding protein (TBP)-related factor 3, TRF3 (also called TBP2), is a vertebrate-specific member of the TBP family that has a conserved carboxy-terminal region and DNA-binding domain virtually identical to that of TBP (ref. 1). TRF3 is highly expressed during embryonic development, and studies in zebrafish and *Xenopus* have shown that it is required for normal embryogenesis[2,3]. Here we show that zebrafish embryos depleted of Trf3 exhibit multiple developmental defects and, in particular, fail to undergo haematopoiesis. Expression profiling for Trf3-dependent genes identified *mespa*, which encodes a transcription factor whose murine orthologue is required for mesoderm specification[4], and chromatin immunoprecipitation verified that Trf3 binds to the *mespa* promoter. Depletion of Mespa resulted in developmental and haematopoietic defects markedly similar to those induced by Trf3 depletion. Injection of *mespa* messenger RNA (mRNA) restored normal development to a Trf3-depleted embryo, indicating *mespa* is the single Trf3 target gene required for zebrafish embryogenesis. Zebrafish embryos depleted of Trf3 or Mespa also failed to express *cdx4*, a *caudal*-related gene required for haematopoiesis. Mespa binds to the *cdx4* promoter, and epistasis analysis revealed an ordered *trf3–mespa–cdx4* pathway. Thus, in zebrafish, commitment of mesoderm to the haematopoietic lineage occurs through a transcription factor pathway initiated by a TBP-related factor.

To analyse the role of TRF3 during embryonic development, we used antisense morpholino oligonucleotides to ablate Trf3, and as a control Tbp, function in zebrafish embryos. Morpholino oligonucleotides were injected into wild-type one-cell stage fertilized embryos, and depletion of Trf3 and Tbp was analysed by immunoblotting at 6 h post-fertilization (h.p.f.), a time at which expression of both proteins was readily detectable (Supplementary Fig. 1a). Immunoblot analysis confirmed that injection of each morpholino oligonucleotide efficiently and specifically depleted its target gene (Supplementary Fig. 1b). Consistent with previous studies[5], Tbp-depleted embryos appeared to initiate gastrulation but failed to progress past 50% epiboly (Supplementary Fig. 2; $n = 122/150$). By contrast, Trf3-depleted embryos appeared to develop normally until the tailbud stage, but by 14 h.p.f. they exhibited delayed development and necrosis compared with siblings injected with a randomized control morpholino oligonucleotide ($n = 166/177$). Inspection of Trf3-depleted embryos at 21 h.p.f. revealed severe necrosis, although head, trunk and tail rudiments were apparent, suggesting that initial antero-posterior patterning was largely unaffected.

To identify Trf3 target genes, we performed expression profiling in *trf3* morpholino-oligonucleotide-treated embryos (at 6 h.p.f.) using a zebrafish oligonucleotide microarray representing about 12,800 genes. As expected, most genes were unaffected by Trf3 depletion (Supplementary Information). Three such representative genes are shown as controls in the reverse transcription–polymerase chain reaction (RT–PCR) experiment in Fig. 1a, which confirms that their expression is dependent upon Tbp but not Trf3. Using a candidate-based approach, we selected genes whose expression was significantly decreased by Trf3 knockdown (Supplementary Table 1) and which had been previously implicated in embryonic development. These candidates were further analysed by RT–PCR for Trf3-dependent expression, chromatin immunoprecipitation (ChIP) for selective Trf3 occupancy and finally for a role in zebrafish development (see below; and data not shown). This combined analysis identified *mespa*, whose mouse orthologue, *Mesp1*, encodes a basic-helix–loop–helix (bHLH)-type transcription factor required for proper embryonic development[4].

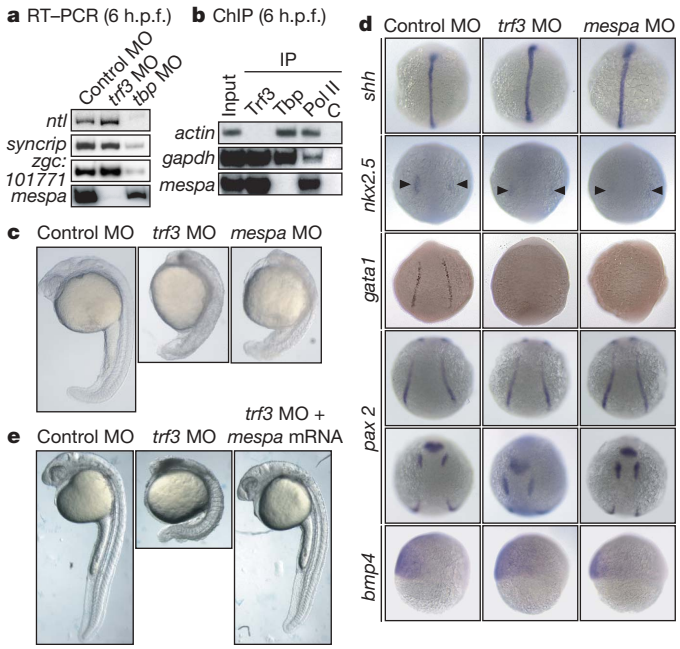


Figure 1 | *mespa* is the single Trf3 target gene required for proper embryonic development. a, RT–PCR analysis. MO, morpholino oligonucleotide. **b**, ChIP analysis. C, negative control (yeast Gal4). **c**, Phenotypic analysis at 24 h.p.f. **d**, Whole-mount *in situ* hybridization with riboprobes to *shh* (14 h.p.f.), *nkx2.5* (10 h.p.f.; arrowheads denote decreased expression in cardiac mesoderm in *trf3* morpholino-oligonucleotide- and *mespa* morpholino-oligonucleotide-treated embryos), *gata1* (12 h.p.f.), *pax2* (10 h.p.f.; showing dorsal posterior (top) and anterior (bottom) views) and *bmp4* (6 h.p.f.). **e**, Phenotypic analysis of embryos injected with a control morpholino oligonucleotide or *trf3* morpholino oligonucleotide, or *trf3* morpholino-oligonucleotide-treated embryos injected with *mespa* mRNA at 28 h.p.f.

[1]Howard Hughes Medical Institute, [2]Programs in Gene Function and Expression and Molecular Medicine, University of Massachusetts Medical School, 364 Plantation Street, Worcester, Massachusetts 01605, USA.

The RT–PCR results (Fig. 1a) show that Trf3 depletion eliminated *mespa* expression. To determine whether *mespa* was a direct Trf3 target, we analysed promoter occupancy by ChIP assays using antibodies directed against Trf3, Tbp, RNA polymerase II (Pol II) or, as a negative control, an irrelevant protein (yeast Gal4). As controls, we analysed two well-characterized housekeeping genes, *actin* and *gapdh*. Figure 1b shows that Pol II was bound to all three promoters, consistent with the transcriptional activity of these genes. Notably, Tbp but not Trf3 was bound to the *actin* promoter, whereas both Tbp and Trf3 were bound to the *gapdh* promoter. By contrast, the *mespa* promoter was selectively bound by Trf3 and not Tbp. Based upon the dependence of *mespa* expression on Trf3 (Fig. 1a) and the selective binding of Trf3 to the *mespa* promoter (Fig. 1b), we conclude that *mespa* is a direct Trf3 target gene.

Phenotypic analysis demonstrated that the *mespa* morpholino-oligonucleotide-injected embryo had a developmental defect that was strikingly similar to that of a Trf3-depleted embryo ($n = 61/89$) (Fig. 1c and Supplementary Fig. 3). As a control, injection of a *mespa* mRNA bearing a silent mutation that prevented hybridization with the *mespa* morpholino oligonucleotide restored normal development to the Mespa-depleted embryo (Supplementary Fig. 4). To compare the phenotypes of the Trf3- and Mespa-depleted embryos in greater detail, we performed whole-mount *in situ* hybridization using several developmentally regulated genes as markers. Figure 1d shows that depletion of either Trf3 or Mespa resulted in increased *shh* expression in axial mesoderm, decreased *nkx2.5* expression in cardiac mesoderm and decreased *gata1* expression in lateral mesoderm. Surprisingly, expression of a second lateral mesoderm marker, *pax2*, was unaffected by loss of Trf3 or Mespa. Moreover, expression of the ventral marker *bmp4* was normal in Trf3- and Mespa-depleted embryos, indicating that the developmental defects were not due to loss of proper dorsal–ventral patterning. Thus the developmental defect observed in Mespa-depleted embryos was very similar, if not identical, to that of Trf3-depleted embryos.

We next asked whether ectopic expression of *mespa* could restore normal development to a Trf3-depleted embryo. Figure 1e shows that injection of *mespa* mRNA ($n = 110/134$), but not an unrelated control mRNA ($n = 0/63$; data not shown), restored normal development to the *trf3* morpholino-oligonucleotide-injected embryo. Completeness of rescue was verified by differential interference contrast microscopy (Supplementary Fig. 5) and *in situ* hybridization analysis (Supplementary Fig. 6). Collectively, these results indicate that in zebrafish, *mespa* is the single Trf3 target gene required for proper embryonic development.

The results shown in Fig. 1d suggested a requirement for Trf3 and Mespa during development of cell types in the lateral mesoderm, which we analysed in greater detail by assaying the expression of several haematopoietic, vascular and pronephric markers in *trf3* or *mespa* morpholino-oligonucleotide-injected embryos. The RT–PCR results shown in Supplementary Fig. 7 indicate that several blood-cell-specific genes were significantly downregulated in Trf3- and Mespa-depleted embryos. These include *hbae1* and *hbae3*, which are terminal markers of erythroid cell fate, as well as *gata1*, which is required for the expression of a variety of genes in the erythroid lineage. In addition, earlier markers of haematopoietic precursors, *scl* and *lmo2*, were similarly reduced although expression of *gata2* was unaffected. Finally, expression of *pu.1*, a marker of myeloid cells that arise in the anterior lateral mesoderm, was also reduced in the absence of Trf3 or Mespa.

To confirm the RT–PCR results, we analysed the expression of these marker genes by whole-mount *in situ* hybridization. Figure 2a shows that in *trf3* or *mespa* morpholino-oligonucleotide-injected embryos, *scl* expression was reduced in the posterior lateral mesoderm, although expression was maintained within more anterior cells whose position is consistent with that of endothelial cells[6,7]. Expression of *lmo2* was moderately reduced in the posterior lateral mesoderm in Trf3- and Mespa-depleted embryos, whereas *pu.1* expression was absent from the anterior lateral mesoderm. Also consistent with the RT–PCR results, expression of the early haematopoietic marker *gata2* was unaffected by loss of Trf3 or Mespa. Similarly, expression of the endothelial cell marker *kdr* was normal in embryos lacking Trf3 or Mespa. As expected, injection of *mespa* mRNA into Trf3-depleted embryos fully rescued expression of the haematopoietic markers *gata1* (see below), *hbae1* and *scl* (Supplementary Fig. 6), as well as the cardiac mesoderm marker *nkx2.5* (Supplementary Fig. 6).

Collectively, the results shown in Fig. 2a suggest a defect in the formation of haematopoietic cells in the posterior lateral mesoderm. To confirm this defect, we simultaneously assayed expression of *fli1*, a marker for both haematopoietic and endothelial cells, and either *gata1* or *pax2*, markers of erythroid or pronephric lateral mesoderm, respectively. Figure 2b shows that expression of a *fli1:EGFP* transgene was maintained in the absence of Trf3 or Mespa, whereas *gata1* expression was absent from the same embryo. By contrast, *pax2* was unaffected by the loss of Trf3 or Mespa and continued to be co-expressed with *fli1* in adjacent cells of the lateral mesoderm. Taken together, these results demonstrate a selective loss of haematopoietic cells within the posterior lateral mesoderm of zebrafish embryos lacking Trf3 or Mespa.

The defects in haematopoietic development described above are reminiscent of those observed in embryos lacking Cdx4, a caudal-related transcription factor that is required for haematopoiesis[6] and which functions by activating expression of homeobox genes involved in the commitment of mesoderm to the haematopoietic lineage[6,8]. RT–PCR and *in situ* hybridization showed that *cdx4* expression was substantially decreased in Trf3-depleted embryos, as well as in Mespa-depleted embryos (Fig. 3a). As expected, expression of the Cdx4-dependent genes *hoxa9a*, *hoxb7a* and *hoxb5a*[8] was also substantially reduced in Trf3- and Mespa-depleted embryos (Supplementary Fig. 8). Depletion of Trf3 or Mespa did not affect expression of *cdx1a* (Supplementary Fig. 9), which has been reported to cooperate with *cdx4* in haematopoietic development[8].

In situ hybridization revealed that at 6 h.p.f. *cdx4*, *mespa* and *trf3* were co-expressed in the presumptive haematopoietic tissues based on the zebrafish fate map[9,10] (Fig. 3b), suggesting that the three factors are components of a common pathway. To confirm this idea and

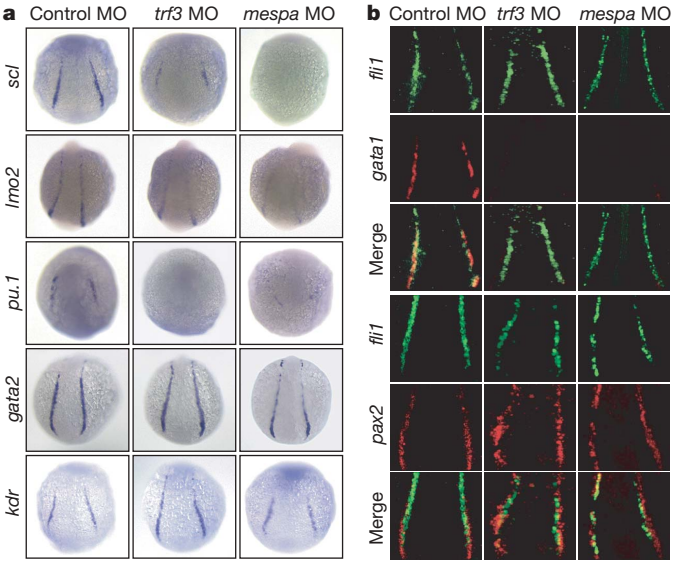


Figure 2 | Trf3- and Mespa-depleted embryos fail to undergo haematopoiesis. a, Whole-mount *in situ* hybridization with riboprobes to *scl* (12 h.p.f.), *lmo2* (14 h.p.f.), *pu.1* (12 h.p.f.), *gata2* (14 h.p.f.) and *kdr* (12 h.p.f.). **b**, Fluorescence microscopy monitoring expression of a *fli1:EGFP* transgene (top panels), and either *gata1* or *pax2* (middle panels). The merged signal is shown in the bottom panels.

to determine the order of the pathway, we performed epistasis experiments using *gata1* expression as a phenotypic read-out. Figure 3c shows that injection of *trf3* mRNA restored *gata1* expression in Trf3-depleted embryos, but not in Mespa- and Cdx4-depleted embryos, indicating that *trf3* is upstream of both *mespa* and *cdx4*. Moreover, injection of *mespa* mRNA restored *gata1* expression in Trf3- and Mespa-depleted embryos, but not in Cdx4-depleted embryos, indicating that *mespa* is upstream of *cdx4* and downstream of *trf3*. Finally, injection of *cdx4* mRNA restored *gata1* expression in Trf3-, Mespa- and Cdx4-depleted embryos, indicating that *cdx4* is downstream of both *trf3* and *mespa*. ChIP analysis supports the possibility that *cdx4* is a direct Mespa target gene (Supplementary Fig. 10a).

Previous studies have shown that ectopic expression of *cdx4* in zebrafish results in increased numbers of blood cells, as evidenced by expanded expression of haematopoietic markers[6]. Ectopic expression of *mespa* also resulted in expanded expression of *gata1* and *scl* (Fig. 3d) and *cdx4* (Fig. 3e). The collective results of Figs 2 and 3 reveal that *trf3*, *mespa* and *cdx4* function in a common pathway, which is summarized in Supplementary Fig. 10b.

Collectively, our results indicate that binding of Trf3 to the *mespa* promoter is the earliest documented step in committing mesoderm to the haematopoietic lineage in the developing zebrafish embryo. A recent study has provided definitive evidence for the existence of hemangioblasts, bipotential progenitors that can give rise to both endothelial and haematopoietic cells, within the zebrafish embryo[10]. Importantly, in zebrafish embryos, hemangioblasts are present at the same time and place in which *trf3*, *mespa* and *cdx4* are expressed. Thus, the timing and location of *trf3*, *mespa* and *cdx4* co-expression suggests that this transcription factor pathway may function within hemangioblasts to specify haematopoietic progenitor cells (that is, commitment of mesoderm to the haematopoietic lineage). Although the defects associated with loss of Trf3 or Mespa are restricted to haematopoietic cell types in the lateral mesoderm, the effects on other tissues appear to be more widespread. We predict that there will be additional Mespa target genes that function analogously to but independently of *cdx4* in mediating other developmental pathways, such as specification of cardiac mesoderm, in the early embryo.

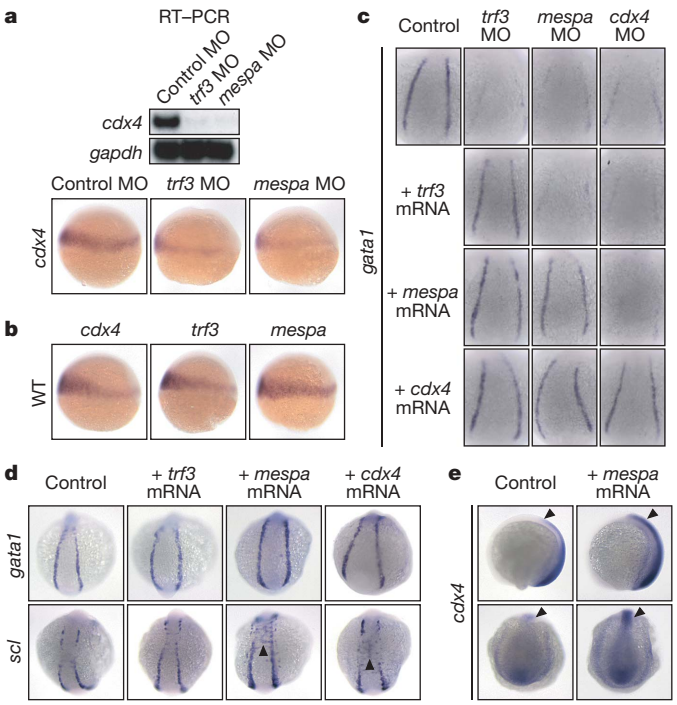


Figure 3 | Trf3 initiates a transcription factor pathway required for *cdx4* expression and haematopoiesis. a, (Top) RT–PCR analysis (9 h.p.f.). (Bottom) *In situ* hybridization (6 h.p.f.). **b**, *In situ* hybridization monitoring *cdx4*, *trf3* and *mespa* expression in wild-type embryos (6 h.p.f.). **c**, Epistasis analysis; *gata1* expression was monitored by *in situ* hybridization at 14 h.p.f. **d**, *In situ* hybridization monitoring *gata1* and *scl* induction (14 h.p.f.). Arrowheads show a 'third stripe' indicative of expanded *scl* expression. **e**, *In situ* hybridization monitoring *cdx4* induction (14 h.p.f.). Lateral (top) and dorsal (bottom) views are shown. Arrowheads indicate normal (left) and expanded (right) *cdx4* expression.

METHODS SUMMARY

Zebrafish maintenance, embryo production and microinjection. Zebrafish (*Danio rerio*) were maintained under standard conditions[11], and embryos were produced and staged as described[12]. Antisense morpholino oligonucleotides were designed against the start codon/5′ untranslated region to block translation. Morpholino oligonucleotide sequences were as follows: control morpholino oligonucleotide, 5′-CCTCTTACCTCAGTTACAATTTATA-3′; *trf3* morpholino oligonucleotide, 5′-GATGCCTCCTCATCCATGTTCAT-3′; and *mespa* morpholino oligonucleotide, 5′-GAAGAGAAAACGTGGAGGCGTCCAT-3′. The *cdx4* morpholino oligonucleotide (5′-CTCCAAAAGGTATCCAACGTACATG-3′) was purchased from Open Biosystems, and the *tbp* morpholino oligonucleotide has been previously described[5]. Morpholino oligonucleotides were injected at a concentration of 5 mg ml^{-1}, except the *cdx4* morpholino oligonucleotide, which was injected at 0.2 mg ml^{-1}. Embryos were analysed morphologically up to 24 h after injection.

To generate mRNAs for phenotypic rescue and epistasis experiments, full-length *trf3*, *mespa* and *cdx4* genes were amplified from shield-stage complementary DNA (cDNA) by RT–PCR (primer sequences are listed in Supplementary Table 2), and subcloned into pCS2 (ref. 13) for mRNA synthesis. The mRNAs were then injected into morpholino-oligonucleotide-treated or wild-type embryos at the one- to two-cell stage at concentrations of 100 ng μl^{-1}, 200 ng μl^{-1} and 30 ng μl^{-1} for *trf3*, *mespa* and *cdx4*, respectively.

Immunoblot analysis. Protein extracts were prepared from zebrafish embryos as described[11]. Blots were probed with either an α-zebrafish Trf3 polyclonal antibody raised to the unique amino-terminal region of zebrafish Trf3 (CPQKESTQADIDTSNS; amino acids 90–105), or an α-human TRF3 polyclonal[1], α-TBP monoclonal (3G3; Eurogentec), or α-RNA Pol II monoclonal (8WG16; Covance Research Products) antibody.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 12 September; accepted 4 October 2007.
Published online 28 November 2007.

1. Persengiev, S. P. *et al.* TRF3, a TATA-box-binding protein-related factor, is vertebrate-specific and widely expressed. *Proc. Natl Acad. Sci. USA* **100,** 14887–14891 (2003).
2. Bartfai, R. *et al.* TBP2, a vertebrate-specific member of the TBP family, is required in embryonic development of zebrafish. *Curr. Biol.* **14,** 593–598 (2004).
3. Jallow, Z., Jacobi, U. G., Weeks, D. L., Dawid, I. B. & Veenstra, G. J. Specialized and redundant roles of TBP and a vertebrate-specific TBP paralog in embryonic gene regulation in *Xenopus*. *Proc. Natl Acad. Sci. USA* **101,** 13525–13530 (2004).
4. Kitajima, S., Takagi, A., Inoue, T. & Saga, Y. MesP1 and MesP2 are essential for the development of cardiac mesoderm. *Development* **127,** 3215–3226 (2000).
5. Muller, F., Lakatos, L., Dantonel, J., Strahle, U. & Tora, L. TBP is not universally required for zygotic RNA polymerase II transcription in zebrafish. *Curr. Biol.* **11,** 282–287 (2001).
6. Davidson, A. J. *et al.* cdx4 mutants fail to specify blood progenitors and can be rescued by multiple hox genes. *Nature* **425,** 300–306 (2003).
7. Sumanas, S. & Lin, S. Ets1-related protein is a key regulator of vasculogenesis in zebrafish. *PLoS Biol.* **4,** e10 (2006).
8. Davidson, A. J. & Zon, L. I. The caudal-related homeobox genes cdx1a and cdx4 act redundantly to regulate hox gene expression and the formation of putative hematopoietic stem cells during zebrafish embryogenesis. *Dev. Biol.* **292,** 506–518 (2006).
9. Langeland, J. & Kimmel, C. B. in *Embryology: Constructing the Organism* (eds Gilbert, S. F. & Raunio, A. M.) 383–407 (Sinauer Associates, Sunderland, Massachusetts, 1997).
10. Vogeli, K. M., Jin, S. W., Martin, G. R. & Stainier, D. Y. A common progenitor for haematopoietic and endothelial lineages in the zebrafish gastrula. *Nature* **443,** 337–339 (2006).
11. Westerfield, M. *The Zebrafish Book. A Guide for the Laboratory Use of Zebrafish* (Danio rerio) (Univ. of Oregon Press, Eugene, 1993).

12. Kimmel, C. B., Ballard, W. W., Kimmel, S. R., Ullmann, B. & Schilling, T. F. Stages of embryonic development of the zebrafish. *Dev. Dyn.* **203,** 253–310 (1995).
13. Turner, D. L. & Weintraub, H. Expression of achaete-scute homolog 3 in *Xenopus* embryos converts ectodermal cells to a neural fate. *Genes Dev.* **8,** 1434–1447 (1994).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Sagerstrom and M. Tsang for reagents, the Kimmel Cancer Center Microarray Core Facility at Thomas Jefferson University for performing the microarray analysis, members of the Lawson laboratory for technical support, and S. Evans for editorial assistance. The Zebrafish International Resource Center is supported by a grant from the National Institutes of Health – National Center for Research Resources (NIH-NCRR). This work was supported in part by a grant from the National Institutes of Health to M.R.G. M.R.G. is an investigator of the Howard Hughes Medical Institute.

Author Contributions D.O.H., N.D.L. and M.R.G. conceived and designed the experiments. D.O.H. performed the experiments, with the assistance of N.D.L., who performed the experiments shown in Supplementary Fig. 2, and T.R., who performed the chromatin immunoprecipitation assays shown in Fig. 1b and Supplementary Fig. 10a. D.O.H., T.R., N.D.L. and M.R.G. analysed the data. D.O.H., N.D.L. and M.R.G. wrote the paper.

Author Information The microarray data have been deposited in the ArrayExpress database at http://www.ebi.ac.uk/arrayexpress under the accession number E-MEXP-1279. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to N.D.L. (nathan.lawson@umassmed.edu) or M.R.G. (michael.green@umassmed.edu).

METHODS

Microarray analysis. Total RNA was prepared from 50 control-morpholino-oligonucleotide- or *trf3* morpholino-oligonucleotide-injected zebrafish embryos 6 h after morpholino oligonucleotide treatment using Trizol reagent (Invitrogen). The KCC 17K Zebrafish Microarray was constructed using a commercially available 17K Zebrafish OligoLibrary consisting of 16,399 oligonucleotides (XEBLIB96; Compugen/Sigma-Genosys), corresponding to about 12,800 genes, based on UniGene clusters. Microarray construction, hybridization and scanning were performed at the Kimmel Cancer Center Microarray Core Facility at Thomas Jefferson University. Chips were scanned by using a Perkin Elmer ScanArray XL5000 Scanner, software version 3.1; images were quantified by PerkinElmer QuantArray Software 3.0. The raw data for all genes induced or repressed by *trf3* morpholino oligonucleotide treatment is provided in the Supplementary Information. Background threshold values were determined for both microarrays and calculated as the mean plus two (for the *trf3* morpholino oligonucleotide microarray) or three (for the control morpholino oligonucleotide microarray) standard deviations of the channel intensities for the five bacterial probes present on each array. The background threshold values were then subtracted from the channel intensity value of each probe set (representing a single gene). The resulting data from the two microarrays were compared to identify Trf3-affected genes, which were classified as those whose probe values satisfied the following condition: greater than 0 in the control microarray (that is, the gene is significantly expressed over background) and no more than 0 in the *trf3* morpholino oligonucleotide microarray (that is, the gene is significantly downregulated, essentially 'off', upon Trf3 depletion); these genes are listed in the 'Analysed data' tab in the Supplementary Data. To determine transcript identity and gene descriptions, we first annotated the gene list by identifying Unigene numbers (Unigene build 104) that corresponded with the GenBank accession number for each oligonucleotide on the array. We subsequently annotated gene names and descriptions through downloadable data sets obtained from the Zebrafish Information Network (ZFIN) and ENSEMBL (Zv7). In each case, we used FileMaker Pro to generate relational databases to annotate gene sets.

RT–PCR. RT–PCR analysis was performed according to standard protocols[14]. Total RNA was prepared using Trizol reagent (Invitrogen) and RT–PCR products were separated by agarose gel electrophoresis and visualized by autoradiography. Primer sequences are listed in Supplementary Table 2.

Chromatin immunoprecipitation. Embryos (about 3,000 at 6 h.p.f.) were washed in 1× PBS three times, suspended in 1× PBS containing 1% formaldehyde (final concentration), transferred to a 10 ml dounce homogenizer and then dounced in one stroke. The reaction mixture was incubated at room temperature for 15 min and quenched by addition of glycine (0.125 M final concentration) for 10 min. The cells were centrifuged at 14,000 r.p.m. (16,100 *g*) for 10 min at 4 °C, and the pellets were washed twice with ice-cold 1× PBS containing protease inhibitors, and resuspended in 5 ml lysis buffer (50 mM HEPES–KOH, pH 7.5, 150 mM NaCl, 1 mM EDTA, 0.1% sodium deoxycholate, 1% Triton X-100, 0.1% SDS, 1 mM PMSF, and protease inhibitor cocktail). The suspension was then sonicated eight times on ice to generate approximately 500 base-pair fragments. The lysates were centrifuged, pre-cleared with protein-A agarose beads (Upstate Biotech), and then divided into 1.5 ml aliquots per immunoprecipitation. Antibodies for immunoprecipitation were as follows: Trf3, Tbp and Pol II antibodies are described in the Methods Summary; the Mespa antibody was raised to the C-terminal portion of the protein (CYQTQNPVQGDFHS; amino acids 191–204); yeast Gal4 (sc577; Santa Cruz). After addition of the antibody, lysates were incubated for 12 h at 4 °C, and then incubated with protein-A agarose beads (50 µl) for 2 h at 4 °C (5% of the lysate was kept as 'input' before the addition of the antibody). The beads were washed twice with lysis buffer without protease inhibitor, twice with the lysis buffer containing 1 M NaCl, and once with LiCl immune-complex wash buffer (50 mM Tris-HCl, pH 8.0, 0.25 M LiCl, 1 mM EDTA, 0.5% N-P40, 0.5% sodium deoxycholate). Finally, the beads were washed three times in TE and pelleted, and chromatin was eluted from the beads by adding 500 µl freshly prepared 0.1 M $NaHCO_3$ and 1% SDS and incubated with rotation for 15 min at room temperature. After addition of 2 µl proteinase K (18.6 ml ml^{-1}), reverse cross-linking was performed for all samples for 6 h at 65 °C. Chromatin was purified from the beads by phenol extraction followed by alcohol precipitation. The input sample was dissolved in 200 µl and the immunoprecipitated samples were dissolved in 40 µl, 2 µl of which was used in the PCR reaction. ChIP PCR primers are listed in Supplementary Table 2.

***In situ* hybridization.** *In situ* hybridization was performed on whole-mount zebrafish embryos as described[15], or on embryos flat-mounted in glycerol. Digoxygenin-labelled antisense probes were synthesized *in vitro* and obtained as follows: *shh*, *pax2*, *cdx4*, *hoxa9a*, *hoxb5a* and *hoxb7a* (C. Sagerstrom, University of Massachusetts Medical School), *bmp4* (M. Tsang, National Institutes of Health), *gata1*, *ephrinb2a* and *kdr* (N.D.L., University of Massachusetts Medical School), *nkx2.5* (The Zebrafish International Resource Center[16]), *scl* (NIH Zebrafish Gene Collection), *lmo2* (IMAGE clone 7433557), *pu.1* (IMAGE clone 6960940) and *gata2* (IMAGE clone 6789690). For the *trf3* and *mespa* probes, N-terminal fragments of either *trf3* or *mespa* were first subcloned into the vector pGEM-T (Promega), and riboprobes were synthesized from the corresponding constructs as T7 or SP6 transcripts using the DIG RNA labelling mix (Roche). Riboprobe primers are listed in Supplementary Table 2.

Differential interference contrast microscopy. High-resolution differential interference contrast microscopy was performed on a Zeiss Axiophot microscope equipped with a Zeiss AxioCam HRC digital camera.

Fluorescence microscopy. Embryos (14 h.p.f.) expressing the *fli1:EGFP* transgene (*TG*(*fli1:EGFP*)^{y1} (ref. 17)) were hybridized with a riboprobe to either *gata1* or *pax2* (described above) for 16 h at 70 °C. Samples were then washed twice in 2× SSC, 0.1% Tween 20 and 50% formamide for 30 min, once in 2× SSC and 0.1% Tween 20 for 15 min, and twice in 0.2× SSC and 0.1% Tween 20 for 30 min, all at 65 °C, and blocked with Western Blocking Reagent (Roche) diluted in PBS for at least 1 h at room temperature. Primary antibody staining was performed in 100 µl of blocking buffer consisting of a 1:200 dilution of anti-GFP rabbit IgG (A11122; Molecular Probes) and a 1:200 dilution of pre-absorbed anti-DIG Fab fragments (Roche) for 2 h at room temperature with gentle shaking. Embryos were washed six times in blocking buffer for 20 min at room temperature. Secondary antibody staining was performed with goat anti-rabbit Alexa Fluor 488 (A11008; Molecular Probes) at a 1:200 dilution in blocking solution for 2 h at room temperature. Embryos were washed six times in blocking buffer for 20 min at room temperature. Fast Red tablets (Roche) were dissolved (one tablet per 2 ml of buffer) in 0.1 M Tris-HCl, pH 8.3 and 0.1% Tween 20 by shaking, and embryos were stained in 1 ml per sample well for 4 h at room temperature. After staining was assessed, embryos were washed six times in PBST for 20 min and stored in 90% glycerol at 4 °C. Embryos were flat-mounted in glycerol and fluorescence microscopy was performed using a Leica TCS SP2 confocal microscope.

14. Beverley, S. M. Enzymatic amplification of RNA by PCR (RT-PCR). In *Current Protocols in Molecular Biology* (eds Ausubel, F. M. *et al.*) Ch. 15.5.1–15.5.6 (John Wiley & Sons, Hoboken, New Jersey, 2001).
15. Hauptmann, G. & Gerster, T. Two-color whole-mount *in situ* hybridization to vertebrate and *Drosophila* embryos. *Trends Genet.* **10,** 266 (1994).
16. Thisse, B. *et al.* Expression of the zebrafish genome during embryogenesis (NIH R01 RR15402). *ZFIN Direct Data Submission.* [online] ⟨http://zebrafish.org/zirc.home/guide.php⟩ (2001).
17. Lawson, N. D. & Weinstein, B. M. *In vivo* imaging of embryonic vascular development using transgenic zebrafish. *Dev. Biol.* **248,** 307–318 (2002).

LETTERS

CLOCK-mediated acetylation of BMAL1 controls circadian function

Jun Hirayama[1], Saurabh Sahar[1], Benedetto Grimaldi[1], Teruya Tamaru[2], Ken Takamatsu[2], Yasukazu Nakahata[1] & Paolo Sassone-Corsi[1]

Regulation of circadian physiology relies on the interplay of interconnected transcriptional–translational feedback loops[1,2]. The CLOCK–BMAL1 complex activates clock-controlled genes, including cryptochromes (*Cry*s), the products of which act as repressors by interacting directly with CLOCK–BMAL1[3,4]. We have demonstrated that CLOCK possesses intrinsic histone acetyltransferase activity and that this enzymatic function contributes to chromatin-remodelling events implicated in circadian control of gene expression[5]. Here we show that CLOCK also acetylates a non-histone substrate: its own partner, BMAL1, is specifically acetylated on a unique, highly conserved Lys 537 residue. BMAL1 undergoes rhythmic acetylation in mouse liver, with a timing that parallels the downregulation of circadian transcription of clock-controlled genes. BMAL1 acetylation facilitates recruitment of CRY1 to CLOCK–BMAL1, thereby promoting transcriptional repression. Importantly, ectopic expression of a K537R-mutated BMAL1 is not able to rescue circadian rhythmicity in a cellular model of peripheral clock. These findings reveal that the enzymatic interplay between two clock core components[6,7] is crucial for the circadian machinery.

The histone acetyltransferase (HAT) function of the master circadian regulator CLOCK[1,5,8–10] indicated that its enzymatic activity could also target non-histone proteins, a feature shown by other HATs[11,12]. In a search for possible targets, we focused on the heterodimerization partner BMAL1, a core clock protein essential for the maintenance of circadian rhythmicity[13,14].

First, we analysed whether various clock proteins, such as BMAL1, CLOCK and PER1, may be acetylated in the mouse liver. Liver extracts were prepared from entrained mice at different zeitgeber times (ZT) (Fig. 1a, left panel). As previously reported, these proteins show rhythmicity in their abundance and phosphorylation levels[15,16]. BMAL1 shows a robust acetylation in the liver that oscillates in a circadian manner, peaking at ZT15 (Fig. 1a, middle panel). In contrast, no acetylation of PER1 and CLOCK could be detected at any time point (Fig. 1a, middle and right panels). Importantly, several additional nuclear proteins and transcription factors were tested and found not to be acetylated (not shown), underscoring the specificity of the assay.

The circadian acetylation of BMAL1 in the liver (Fig. 1a) and our recent finding of CLOCK's HAT activity[5] indicated that CLOCK could mediate BMAL1 acetylation. Thus, we expressed Flag–Myc–BMAL1 in the presence or absence of Myc–CLOCK in cultured mammalian cells. The results unequivocally revealed that CLOCK induces BMAL1 acetylation (Fig. 1b), whereas CLOCK and PER1 do not undergo acetylation (Supplementary Fig. 1), emphasizing the specificity of CLOCK-mediated enzymatic function. To establish whether heterodimerization is required for acetylation, we used an amino-terminally truncated CLOCK (CLOCKΔN) mutant that lacks the PAS domains and hence is unable to interact with BMAL1. Importantly, CLOCKΔN still possesses HAT activity[5]. CLOCKΔN did not induce BMAL1 acetylation, indicating that CLOCK–BMAL1 dimerization is essential for CLOCK-dependent acetylation of BMAL1 (Fig. 1c). We then used a CLOCK protein with three single amino acid mutations in the HAT domain (CLOCK(mutA)), which we had previously shown to have reduced HAT activity[5]. BMAL1 acetylation by CLOCK(mutA) was drastically reduced compared to wild-type CLOCK (CLOCK(WT)), demonstrating that intrinsic HAT activity of CLOCK is required for BMAL1 acetylation (Fig. 1c). Next, we addressed the possibility that negative circadian regulators[17–19] may influence CLOCK-dependent acetylation of BMAL1. We found that neither CRY1 nor PER2 affect BMAL1 acetylation (Supplementary Fig. 2).

We have reported that BMAL1 is SUMOylated[20]. Four candidate lysines—K223, K229, K259 and K272—were identified as SUMOylatable in mouse BMAL1, K259 being the major *in vivo* SUMOylation site[20]. Because SUMOylatable lysines could also be subject to acetylation[21], we tested whether these could be targets for CLOCK-mediated acetylation. Site-directed mutagenesis of these residues, either alone or in combination, demonstrated that none of them is acetylated by CLOCK (Supplementary Fig. 3). Thus, the target lysines for the two modifications are distinct.

To identify the lysine(s) acetylated by CLOCK in BMAL1, we first generated two carboxy-terminally truncated BMAL1 proteins (amino acids 1–282 and 1–469) (Fig. 2a). Although both mutant proteins interact with CLOCK (Supplementary Fig. 4), they were not acetylated, indicating that the acetylatable lysine(s) must be located in the C terminus (amino acids 470–631 of mouse BMAL1; Fig. 2b). This region contains four potential target lysines at positions 475, 494, 537 and 538, which were individually mutated into arginine by site-directed mutagenesis. All mutant proteins are acetylated at levels comparable to wild-type BMAL1, with the exception of K537R (Fig. 2c). Importantly, this lysine is highly conserved among all vertebrate BMAL1s (Fig. 2a).

To establish unequivocally that BMAL1 is indeed acetylated by CLOCK, we used recombinant BMAL1 protein in an *in vitro* acetylation assay. Anti-Myc antibodies were used to isolate either Myc–CLOCK or Myc–GFP (green fluorescent protein) after ectopic expression in cultured cells (Fig. 2e). Myc–CLOCK or Myc–GFP were incubated with a bacterially purified glutathione *S*-transferase (GST)–BMAL1 in the presence of ^{3}H-acetyl-coA, and acetylation was analysed by SDS–PAGE followed by autoradiography. CLOCK was able to acetylate recombinant BMAL1, whereas GFP did not. Notably, recombinant BMAL1(K537R) was not acetylated by CLOCK (Fig. 2d), confirming the target specificity of CLOCK-mediated acetylation (Fig. 2c).

[1]Department of Pharmacology, School of Medicine, University of California, Irvine, 92697-4625 Irvine, California, USA. [2]Department of Physiology, Toho University, Faculty of Medicine, Tokyo 143-8540, Japan.

Next, we addressed the physiological relevance of BMAL1 acetylation for circadian rhythmicity by performing rescue experiments using mouse embryonic fibroblasts (MEFs) generated from *Bmal1*$^{-/-}$ mice. As described previously, lack of BMAL1 in MEFs results in a dysfunctional circadian clock and arrhythmic gene expression[20,22]. We infected *Bmal1*$^{-/-}$ MEFs with retroviruses expressing either wild-type BMAL1, an acetylation-deficient BMAL1 (K537R) or GFP (Fig. 3a). Importantly, only one additional copy of the *Bmal1* gene is transduced per cell by retroviral infection[20]. Also, Myc–BMAL1 expression is under control of the *Bmal1* promoter to mimic the natural regulation of the gene. Infected MEFs

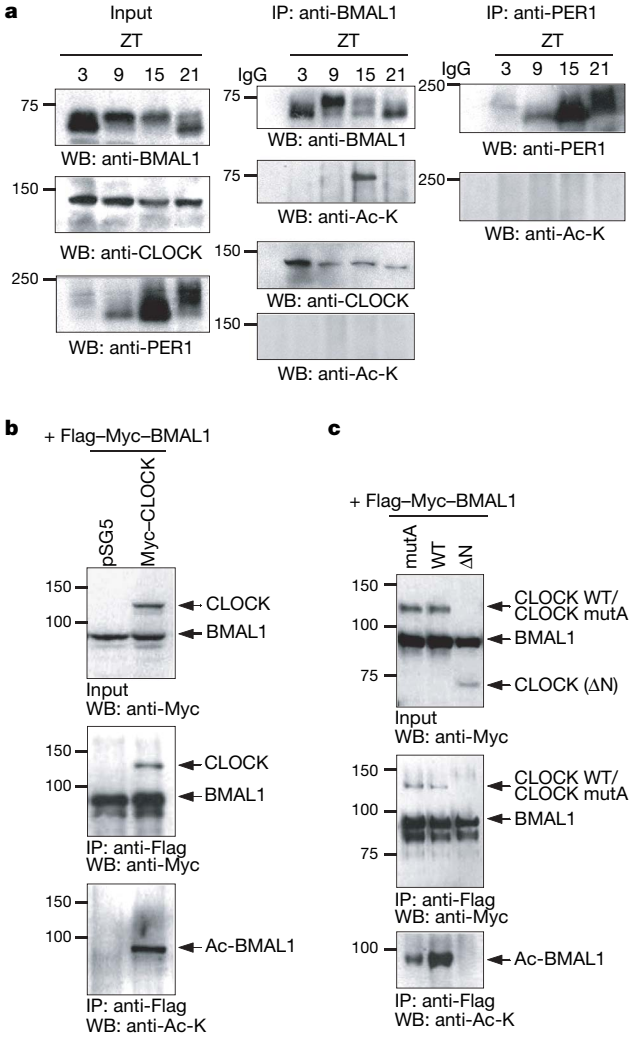


Figure 1 | Circadian regulator BMAL1 is acetylated. **a**, Circadian acetylation of BMAL1 in mouse liver. Left panels, equal protein amounts from mouse liver extracts at the indicated zeitgeber times (ZT) were immunoblotted with antibodies against mouse BMAL1, CLOCK or PER1. Middle panels, mouse BMAL1 was immunoprecipitated from each lysate with anti-BMAL1 antibody. Immunoprecipitates were immunoblotted with either anti-BMAL1, anti-pan-acetyl lysine or anti-CLOCK antibodies. Right panels, mouse PER1 was immunoprecipitated from each lysate with anti-PER1 antibody, and the immunoprecipitates were immunoblotted with either anti-PER1 antibody or anti-pan-acetyl lysine antibody. **b**, CLOCK induces BMAL1 acetylation. BMAL1 was immunoprecipitated using anti-Flag antibody from lysates of JEG3 cells transiently expressing Flag–Myc–BMAL1 alone (left lane) or both Flag–Myc–BMAL1 and Myc–CLOCK (right lane). Top panel, immunoblot analysis of total cell lysates with anti-Myc antibody. Middle panel, immunoblot analysis of precipitated Flag–Myc–BMAL1 and co-precipitated Myc–CLOCK, detected with anti-Myc antibody. Bottom, immunoblot analysis of precipitated acetylated BMAL1 (Ac-BMAL1), detected with anti-pan-acetyl lysine antibody. **c**, Both CLOCK–BMAL1 dimerization and CLOCK intrinsic HAT activity are required for CLOCK-induced BMAL1 acetylation. JEG3 cells were co-transfected with Flag–Myc–BMAL1 and Myc–CLOCK(mut A) (lane 1), Myc–CLOCK(WT) (lane 2), or Myc–CLOCK(ΔN) (lane 3). Flag–Myc–BMAL1 was immunoprecipitated using Flag antibody and precipitates analysed by immunoblotting as described in **b**. All protein sizes are given in kDa.

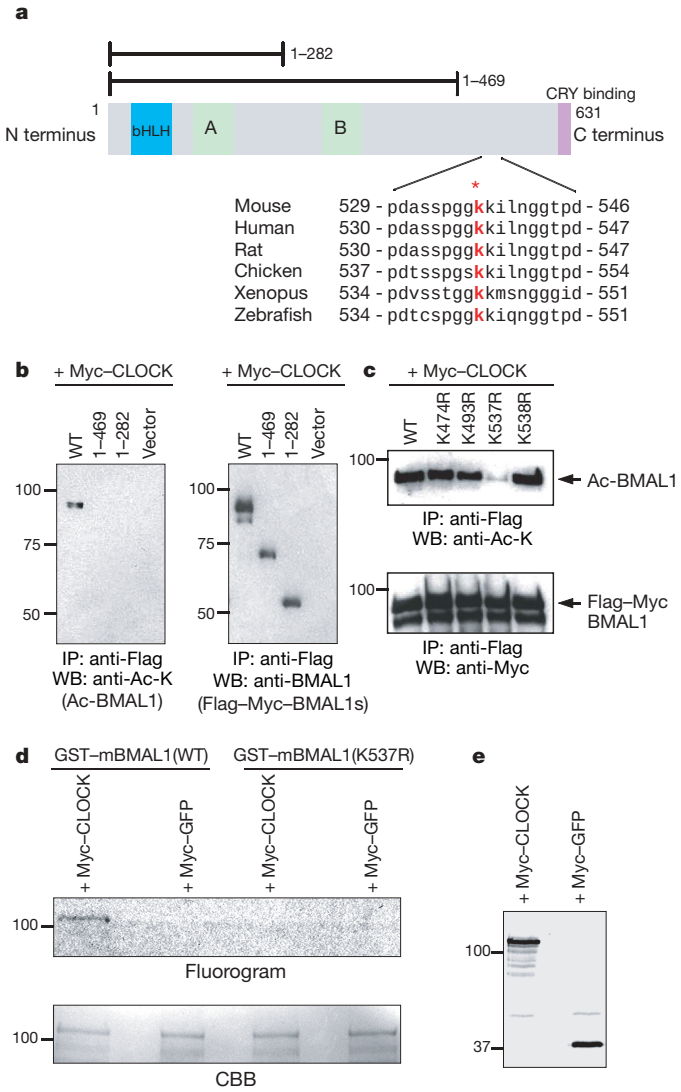


Figure 2 | A single lysine of BMAL1 is acetylated. **a**, Schematic representation of mouse BMAL1 protein showing the positions of the basic helix–loop–helix (bHLH), PAS (PER-ARNT-SIM) A (A), PAS B (B) and CRY-binding site (CRY binding)[29]. Numbers indicate amino acid residues in the mouse protein. The extents of deletion mutants (amino acids 1–282 and 1–469) are shown as grey bars on top. Sequence alignment of BMAL1 with the target lysine for acetylation from various species is shown; the target lysines are highlighted in red. **b**, **c**, Identification of the target lysine for acetylation. Cells were co-transfected with expression vectors for Myc–CLOCK and Flag–Myc–BMAL1 wild type or each BMAL1 mutant. Acetylation of immunoprecipitated Flag–Myc–BMAL1 was determined by western blotting using anti-pan-acetyl-Lys (left panel in **b** and upper panel in **c**). Expression of precipitated BMAL1 was determined by western blotting using anti-BMAL1 (**b**, right panel) or anti-Myc (**c**, bottom panel) antibodies. **d**, *In vitro* assay of CLOCK-induced BMAL1 acetylation. Myc–CLOCK or Myc–GFP expressed in JEG3 cells were immunoprecipitated using anti-Myc antibody. The resulting immunoprecipitates were incubated with bacterially expressed GST–BMAL1(WT) or –BMAL1(K537R) in the presence of ^{3}H-acetyl-coA. Reactions were resolved by SDS–PAGE, gels stained with Coomassie brilliant blue (CBB) and analysed by fluorography. **e**, Immunoblot analysis of precipitated Myc–CLOCK and Myc–GFP, detected with anti-Myc antibody.

were synchronized by dexamethasone (Dex) treatment and circadian oscillation was monitored by real-time bioluminescence using a *Per2*-promoter-driven luciferase reporter vector[23,24]. Infection of *Bmal1*$^{-/-}$ MEFs with a virus expressing wild-type BMAL1 rescued circadian *Per2* expression, whereas the BMAL1(K537R) mutant was unable to do so. Thus, BMAL1 acetylation is essential for circadian regulation of gene expression (Fig. 3b).

The drastic effect of BMAL1 acetylation on circadian rhythmicity prompted us to investigate the underlying molecular mechanism. Importantly, the lack of rescue by BMAL1(K537R) is not due to impaired recruiting to the *Per2* promoter, as demonstrated by chromatin immunoprecipitation (ChIP) assays using anti-Myc antibody in the *Bmal1*$^{-/-}$ cells rescued with either Myc–BMAL1 or Myc–BMAL1(K537R). Indeed, both BMAL1 and the acetylation-deficient mutant are recruited to the promoter with equivalent efficiency (Supplementary Fig. 5). Moreover, the K537R mutation has no effect on protein stability, subcellular localization and CLOCK-induced phosphorylation of BMAL1 (Supplementary Figs 6 and 7). Finally, the association capacity of BMAL1(K537R) with CLOCK is essentially equivalent to wild-type BMAL1, as demonstrated by mammalian two-hybrid (Supplementary Fig. 8a) and co-immunoprecipitation assays (Fig. 4a).

We reasoned that BMAL1 acetylation may be involved in modulating CRY1-mediated repression. This possibility would rationalize the results of the circadian rescue experiments (Fig. 3b) and is supported by the notion that impairment of CRY1-mediated repression of CLOCK–BMAL1 leads to loss of circadian rhythmicity[24]. In fact, the BMAL1(K537R) mutant showed a drastically reduced sensitivity to CRY1-mediated repression compared to wild-type BMAL1 (Fig. 4b). As a consequence, we predict that *Per* and *Cry* expression should be upregulated in cells expressing the BMAL1(K537R) mutant compared to cells in which circadian rescue was successfully achieved by expressing wild-type BMAL1. This is indeed what we found by analysing both RNA and protein levels (Supplementary Fig. 9). In other words, acetylation of BMAL1 by CLOCK is an essential regulatory switch because it facilitates CRY1-dependent repression. Finally, we used a BMAL1(K538R) mutant, which carries a Lys→Arg substitution at the residue adjacent to the CLOCK-target Lys 537 (Fig. 2a). The BMAL1(K538R) mutant shows sensitivity to CRY1-mediated repression, analogous to wild-type BMAL1 (Fig. 4b). This result validates the physiological importance of Lys 537 acetylation and stresses its remarkable specificity. Hence, our data legitimize the complete loss of circadian rhythmicity caused by the K537R mutation (Fig. 3b).

To repress transcription, CRY1 needs to bind directly to the CLOCK–BMAL1 complex[25,26], and our transactivation studies indicate that BMAL1 acetylation could be a mark for CRY recruitment (Fig. 4b). Thus, we examined whether K537 acetylation is required for efficient recognition of the CLOCK–BMAL1 complex by CRY1. First, we performed a mammalian two-hybrid assay, in which CRY1 fused to the GAL4 DNA-binding domain (GAL4-CRY1) was co-expressed with BMAL1 fused to the VP16 transactivation domain (VP16-BMAL1) and CLOCK. When VP16-BMAL1 interacts functionally with GAL4-CRY1, VP16 is recruited to the vicinity of the promoter and elicits transactivation. Importantly, activation induced by VP16-BMAL1(K537R) was dramatically reduced compared to that by VP16-BMAL1(WT) (Supplementary Fig. 8b). The reduced interaction between CRY1 and BMAL1(K537R) was also confirmed by co-immunoprecipitation assays (Fig. 4c). Notably, the BMAL1 (K538R) mutant, which has the same sensitivity to CRY1-mediated repression as wild-type BMAL1 (Fig. 4b), showed unaltered interaction ability both in two-hybrid and co-immunoprecipitation assays (Supplementary Figs 8b, 4c).

Taken together, these data demonstrate that BMAL1 acetylation facilitates CRY1 interaction with the CLOCK–BMAL1 complex.

Our findings demonstrate that CLOCK exerts its enzymatic activity also on non-histone proteins, and more importantly on BMAL1, a core component of the clock machinery. Importantly, BMAL1 acetylation occurs *in vivo* and is regulated in a circadian manner (Fig. 1a). As BMAL1 enhances the intrinsic HAT activity of CLOCK[5], it is tempting to speculate that BMAL1 modulates its own acetylation by reciprocally controlling CLOCK enzymatic activity. Another possibility is that NPAS2, an alternative partner of BMAL1, could also operate in modulating acetylation of BMAL1 and/or the enzymatic function of CLOCK. Because NPAS2 and CLOCK have differential cellular distribution, it would be of interest to establish in what manner these two proteins may compensate for each other's activity. As a functional CLOCK protein is specifically required for peripheral oscillators[27], it is possible that the regulation described here could be part of a tissue-specific circadian pathway of regulation.

Acetylation of proteins is an essential regulatory mechanism, having both stimulatory and inhibitory effects on transcription[28]. Here we have demonstrated that acetylation may operate at yet another level of control, because BMAL1 acetylation serves to increase the repressive function of another regulator, CRY1. Our results also indicate that CLOCK enzymatic activity has a dual regulatory

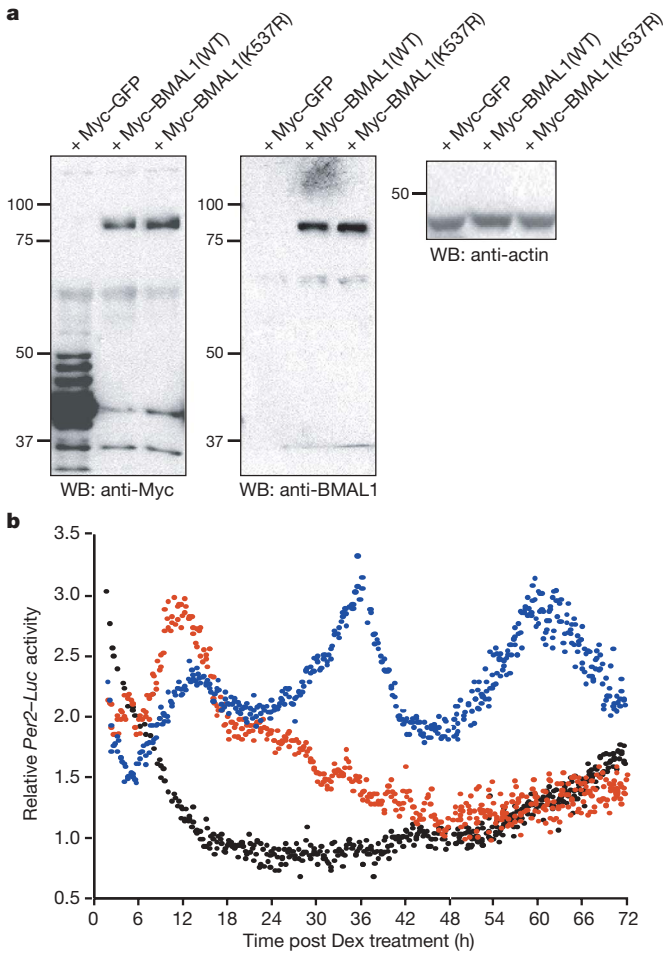


Figure 3 | Acetylation of BMAL1 is essential to rescue the circadian rhythmicity in BMAL1-deficient cells. a, Expression levels of BMAL1 and GFP proteins in retrovirus-infected *Bmal1*$^{-/-}$ MEFs were evaluated using anti-Myc (left panel) and anti-BMAL1 immunoblots (middle panel). The same cell lysates were immunoblotted with an antibody against actin as a loading control (right panel). **b**, A mouse *Per2*-promoter luciferase reporter plasmid[24] was transfected into the retrovirus-infected cells and *Per2* promoter activity was monitored by a real-time bioluminescence assay. Expression levels were plotted as arbitrary units. Results are representative of three independent experiments. Blue, red and black plots indicate the results from MEFs expressing BMAL1(WT), BMAL1(K537R) and GFP, respectively.

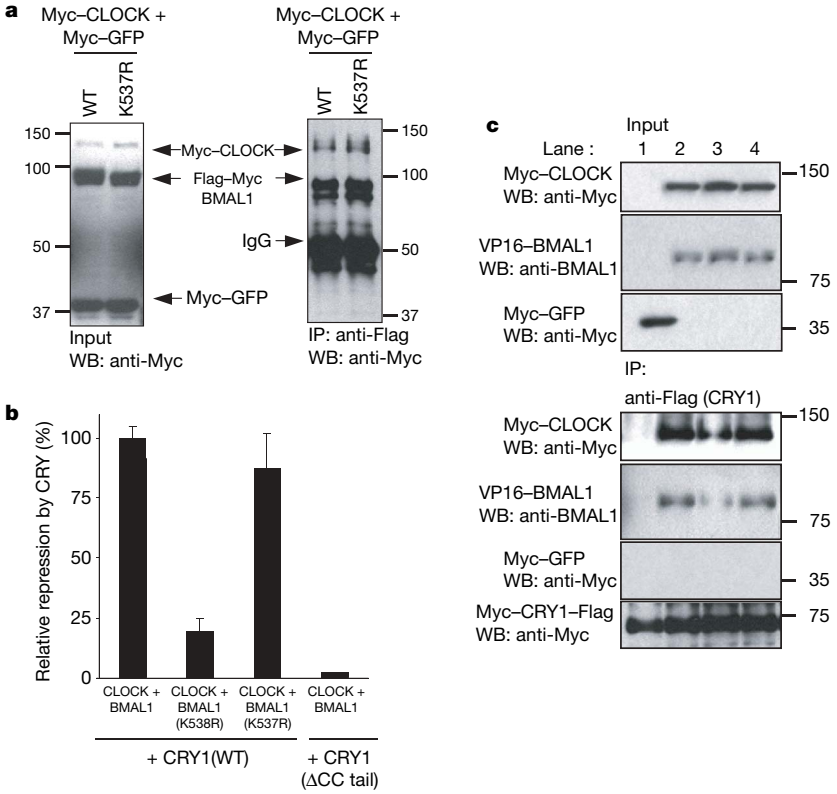


Figure 4 | Acetylation of BMAL1 facilitates CRY1-mediated repression. **a**, Heterodimerization of CLOCK and BMAL1 is not influenced by acetylation at K537. CLOCK and BMAL1 interaction was tested by co-immunoprecipitation. Expression vectors for Myc–CLOCK (700 ng), Myc–GFP (400 ng), and Flag–Myc–BMAL1(WT) (700 ng) or Flag–Myc–BMAL1(K537R) (700 ng) were co-transfected in JEG3 cells. Cell lysates were immunoprecipitated with anti-Flag antibody. **b**, Repression mediated by CRY1 on CLOCK–BMAL1 is impaired by lack of acetylation at K537. Repression by CRY1 on CLOCK–BMAL1-mediated activation of a *Per2* reporter promoter in a luciferase assay. Here 1 ng of the CRY1 expression vector was used and repression on CLOCK–BMAL1 is indicated as 100%. Repression on the CLOCK–BMAL1(K537R) dimer is reduced by fivefold (error bars, s.e.m.). As a control for lack of repression we used the CRY(ΔCCtail) mutant[26] on the CLOCK–BMAL1 dimer. Analogous results were obtained by using a large range of expression vectors for CRY1, CLOCK and BMAL1. In all assays, CRY1-mediated repression on the CLOCK–BMAL1(K537R) dimer was reduced with respect to CLOCK–BMAL1 by 3–5-fold. **c**, BMAL1 acetylation facilitates association of CRY1 with CLOCK–BMAL1. CLOCK–BMAL1 interaction with CRY1 was tested by co-immunoprecipitation. Expression vectors for Myc–CRY1–Flag (300 ng) and Myc–GFP (2,000 ng) (lane1), Myc–CRY1–Flag (100 ng), Myc–CLOCK (3,000 ng) and VP16-BMAL1(WT) (3,000 ng) (lane 2), Myc–CRY1–Flag (100 ng), Myc–CLOCK (3,000 ng), and VP16-BMAL1(K537R) (3,000 ng) (lane 3), or Myc–CRY1–Flag (100 ng), Myc–CLOCK (3,000 ng), and VP16-BMAL1(K538R) (3,000 ng) (lane 4) were co-transfected in JEG3 cells. Lysates were immunoprecipitated with anti-Flag antibody. Immunoprecipitates and total cell lysates were analysed by immunoblotting with anti-BMAL1 antibody (BMAL1) or anti-Myc antibody (CLOCK, GFP and CRY1).

function. Here we have shown that it contributes to the negative limb of the circadian feedback loop, whereas CLOCK-mediated acetylation of histones[5] participates in the transcriptional stimulation of clock-controlled genes, acting within the positive limb of the loop. Thus, CLOCK enzymatic function contributes in multiple ways to the time-dependent regulation of circadian physiology.

METHODS SUMMARY

Acetylation of BMAL1 in liver extracts was assessed after immunoprecipitation with an anti-BMAL1 specific antibody[20] and then by using either polyclonal or monoclonal anti-pan-acetyl-lysine antibodies. Livers were dissected from mice entrained in 12 h light:12 h dark for 14 days before placement in constant darkness. At selected times on the first cycle in constant darkness, animals were killed, and tissues collected and frozen in dry ice. Experimental chronology is measured in ZT, subjective day beginning at 07:00 (ZT0), and subjective night beginning at 19:00 (ZT12). Ectopic expression of clock proteins in cultured cells was obtained by transient transfection of specifically designed expression vectors, as already described[5]. MEFs were prepared and cultured as routinely done in our laboratory[5,20], and were derived from either wild-type mice or *Bmal1*-null mutant mice[7]. MEFs were infected with retrovirus-based vectors expressing either the wild-type BMAL1 protein or BMAL1 mutated proteins in specific lysine residues. To score for circadian gene expression we used a *Per2* promoter luciferase reporter plasmid[24], and promoter activity was monitored by a real-time bioluminescence assay, as described[5]. Mutagenesis, *in vitro* acetylation assays, preparation of protein extracts, association assays and chromatin-immunoprecipitation experiments were performed as described[5,20]. Additional experimental procedures, detailed protocols, sample preparations, microscopy techniques and processing of the data are described in Methods.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 15 July; accepted 16 October 2007.

1. Dunlap, J. C. Molecular bases for circadian clocks. *Cell* **96**, 271–290 (1999).
2. Reppert, S. M. & Weaver, D. R. Coordination of circadian timing in mammals. *Nature* **418**, 935–941 (2002).
3. King, D. P. & Takahashi, J. S. Molecular genetics of circadian rhythms in mammals. *Annu. Rev. Neurosci.* **23**, 713–742 (2000).
4. Young, M. W. & Kay, S. A. Time zones: a comparative genetics of circadian clocks. *Nature Rev. Genet.* **2**, 702–715 (2001).
5. Doi, M., Hirayama, J. & Sassone-Corsi, P. Circadian regulator CLOCK is a histone acetyltransferase. *Cell* **125**, 497–508 (2006).
6. King, D. P. *et al.* Positional cloning of the mouse circadian clock gene. *Cell* **89**, 641–653 (1997).
7. Bunger, M. K. *et al.* Mop3 is an essential component of the master circadian pacemaker in mammals. *Cell* **103**, 1009–1017 (2000).
8. Cermakian, N. & Sassone-Corsi, P. Multilevel regulation of the circadian clock. *Nature Rev. Mol. Cell Biol.* **1**, 59–67 (2000).

9. Hirayama, J. & Sassone-Corsi, P. Structural and functional features of transcription factors controlling the circadian clock. *Curr. Opin. Genet. Dev.* **15,** 548–556 (2005).
10. Belden, W. J., Loros, J. J. & Dunlap, J. C. CLOCK leaves its mark on histones. *Trends Biochem. Sci.* 31, 610– 613 (2006).
11. Glozak, M. A., Sengupta, N., Zhang, X. & Seto, E. Acetylation and deacetylation of non-histone proteins. *Gene* **363,** 15–23 (2005).
12. Zhang, K. & Dent, S. Y. Histone modifying enzymes and cancer: going beyond histones. *J. Cell. Biochem.* **96,** 1137–1148 (2005).
13. Gekakis, N. *et al.* Role of the CLOCK protein in the mammalian circadian mechanism. *Science* **280,** 1564–1569 (1998).
14. Hogenesch, J. B. *et al.* The basic helix-loop-helix-PAS protein MOP9 is a brain-specific heterodimeric partner of circadian and hypoxia factors. *J. Neurosci.* **20,** RC83 (2000).
15. Lee, C., Etchegaray, J. P., Cagampang, F. R., Loudon, A. S. & Reppert, S. M. Posttranslational mechanisms regulate the mammalian circadian clock. *Cell* **107,** 855–867 (2001).
16. Matsuo, T. *et al.* Control mechanism of the circadian clock for timing of cell division *in vivo*. *Science* **302,** 255–259 (2003).
17. van der Horst, G. T. *et al.* Mammalian Cry1 and Cry2 are essential for maintenance of circadian rhythms. *Nature* **398,** 627–630 (1999).
18. Kume, K. *et al.* mCRY1 and mCRY2 are essential components of the negative limb of the circadian clock feedback loop. *Cell* **98,** 193–205 (1999).
19. Jin, X. *et al.* A molecular mechanism regulating rhythmic output from the suprachiasmatic circadian clock. *Cell* **96,** 57–68 (1999).
20. Cardone, L. *et al.* Circadian clock control by SUMOylation of BMAL1. *Science* **309,** 1390–1394 (2005).
21. Shalizi, A. *et al.* A calcium-regulated MEF2 sumoylation switch controls postsynaptic differentiation. *Science* **311,** 1012–1017 (2006).
22. Kondratov, R. V. *et al.* BMAL1-dependent circadian oscillation of nuclear CLOCK: posttranslational events induced by dimerization of transcriptional activators of the mammalian clock system. *Genes Dev.* **17,** 1921–1932 (2003).
23. Nagoshi, E. *et al.* Circadian gene expression in individual fibroblasts: cell-autonomous and self-sustained oscillators pass time to daughter cells. *Cell* **119,** 693–705 (2004).
24. Sato, T. K. *et al.* Feedback repression is required for mammalian circadian clock function. *Nature Genet.* **38,** 312–319 (2006).
25. Griffin, E. A. Jr, Staknis, D. & Weitz, C. J. Light-independent role of CRY1 and CRY2 in the mammalian circadian clock. *Science* **286,** 768–771 (1999).
26. Chaves, I. *et al.* Functional evolution of the photolyase/cryptochrome protein family: importance of the C terminus of mammalian CRY1 for circadian core oscillator performance. *Mol. Cell. Biol.* **26,** 1743–1753 (2006).
27. DeBruyne, J. P., Weaver, D. R. & Reppert, S. M. Peripheral circadian oscillators require CLOCK. *Curr. Biol.* **17,** R538–R539 (2007).
28. Sterner, D. E. & Berger, S. L. Acetylation of histones and transcription-related factors. *Microbiol. Mol. Biol. Rev.* **64,** 435–459 (2000).
29. Kiyohara, Y. B. *et al.* The BMAL1 C terminus regulates the circadian transcription feedback loop. *Proc. Natl Acad. Sci. USA* **103,** 10074–10079 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. S. Steffan, C. A. Bradfield, G. T. van der Horst, F. Tamanini, M. Doi, T. Takumi and T. Todo for discussions and sharing of reagents. We also thank M. Kaluzova, D. Gauthier, D. Mishra Prasad and all colleagues in the Sassone-Corsi laboratory for discussions and help. This work was supported by grants from the Cancer Research Coordinating Committee of the University of California and from the National Institutes of Health to P.S.-C.

Author Contributions J.H., S.S., B.G. and P.S.-C. designed the research; J.H., S.S., B.G., T.T., K.T. and Y.N. performed the experiments; J.H., S.S., B.G., T.T. and P.S.-C. analysed the data; and J.H. and P.S.-C. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to P.S.-C. (psc@uci.edu).

METHODS

Plasmids. Flag–Myc–BMAL1/pCS2 was made by adding the Flag epitope into Myc–BMAL1/pCS2.

A HindIII–EcoRI fragment of Myc epitope was inserted in the corresponding sites of pEGFPN2, generating a Myc–GFP-expressing vector. A HindIII–XhoI fragment of full-length BMAL1 was inserted in the corresponding sites of VP16-pcDNA, generating VP16-BMAL1-pcDNA3. Flag–Myc–BMAL1 mutants were created by using QuickChange site-directed mutagenesis kit (Stratagene). CRY1 expression vectors were a gift of T. Takumi. Other plasmids used in this study have been described elsewhere[5,20].

Antibodies and immunoblotting. Anti-haemagglutinin, anti-Myc, anti-BMAL1, and anti-PER1 antibodies were previously described[5,20]. Monoclonal anti-pan-acetyl-Lysine (Cell Signalling), polyclonal anti-pan-acetyl-Lysine (Cell Signalling), anti-CLOCK (Calbiochem-Novabiochem) and anti-Flag (Sigma) antibodies were purchased. Monoclonal and polyclonal anti-pan-acetyl-lysine (Cell Signalling) antibodies were used for liver extract sample and the other experiments, respectively.

Mice, cell culture, transfections and Dex treatment. Wild-type mice were entrained in 12 h light:12 h dark for 14 days before placement in constant darkness. JEG3 cells were cultured in basal medium Eagle (BME) containing 10% fetal bovine serum (FBS) and transfected with Fugene (Roche), according to the manufacturer's protocol. 293 cells were cultured in Dulbecco's modified Eagle's medium (DMEM) containing 10% newborn calf bovine serum (NCS). MEF cells were cultured in DMEM containing 10% FBS. MEFs from *Bmal1*$^{-/-}$ mice were a gift of C. Bradfield. For luciferase assays, cells growing in 24-well plates were transfected with various combinations of expression plasmids. The total amount of DNA applied per well was adjusted by adding pSG5 vector. Cell extracts were subjected to luminometry-based-luciferase assay (Promega), and luciferase activity was normalized by β-galactosidase activity. All experiments were repeated at least three times. For dexamethazone (Dex) treatment, BMAL1-deficient MEFs infected with *Bmal1*-promoter-driven BMAL1(wild type), BMAL1(K537R) and CMV-promoter-driven GFP were transfected with pGL4-Per2 promoter (~1.8 kb) reporter, using Fugene HD (Roche). Twenty-four hours after the transfection, cells were treated with 0.1 mM Dex for 20 min and then the culture medium was changed to DMEM containing 20% FBS. Real-time luciferase activities were monitored using Kronos (ATTO).

Protein extracts, immunoprecipitations and western analyses. JEG3 cells were cultured with deacetylase inhibitors before they were harvested. JEG3 cells were then lysed in binding buffer (150 mM NaCl, 5 mM EDTA, 0.5% NP-40 and 50 mM Tris-HCl, pH 7.8, 1× protease inhibitor, deacetylase inhibitors, and 1 mM PMSF) or modified radio-immunoprecipitation assay buffer (RIPA-1; 50 mM Tris-HCl, pH 7.8, 150 mM NaCl, 1 mM PMSF, 5mM EDTA, 15 mM $MgCl_2$, 1% Nonidet P-40, 0.5% sodium deoxycholate, 1 mM dithiothreitol, 1× protease inhibitor, deacetylase inhibitors and 1 mM PMSF) for co-immunoprecipitation or the other experiments, respectively. Liver samples (1 g) were manually homogenated by pestle in cold PBS containing deacetylase inhibitors and 1× protease inhibitor cocktail. Pellets of liver cells were washed with the same buffer, then lysed with RIPA-1. Immunoprecipitation experiments were carried out in the RIPA-1 buffer with Flag, HA, BMAL1 or PER1 antibodies. Western analyses were performed as described. Protein samples were resolved on 7% or 6% SDS–PAGE and immunoblotted with the following antibodies: anti-Myc 9E10, anti-HA, anti-BMAL1, anti-PER1, anti-CLOCK, anti-pan-acetyl-lysine and anti-actin.

Immunofluorescence. 293 cells were seeded on glass coverslips in 6-well dishes and transfected the following day with 1 μg of total DNA per well. Thirty hours after transfection, the cells were washed twice with PBS, and fixed with 4% paraformaldehyde in PBS. After several washes, the cell nuclei were stained with 4′,6′-diamidino-2-phenylindole, and the cells were mounted for fluorescence microscopy.

Retroviral infection of MEF cells. RetroMax expression system (IMGENEX) was used to produce retrovirus according to the manufacturer's instructions. Briefly, pCLNCX retroviral vector containing either wild-type or mutated mouse *Bmal1b* genes and enveloping vector, pMD.G/vsv-g, were transfected into HEK293 gag/pol packaging cells. Forty-eight hours after transfection, all the medium containing viral particles was recovered, added to 1 ml FCS and 1 ml polybrene (0.04 mg ml^{-1}), and filtered (viral solution). For infection, MEF cells were incubated with the viral solution for 4 h and this procedure was repeated 4 times. After final incubation, cells were cultured with complete medium and used for the subsequent analysis.

Preparation of GST–BMAL1 proteins. The GST fusion vectors (pGEX-BMAL1(WT) or pGEX-BMAL1(K537R)) were transformed in *Escherichia coli* BL21(DE3). Cells were grown at 25 °C to D_{600} = 0.5–0.8 by induction with 0.1 mM isopropyl-1-thio-β-D-galactopyranoside for 12 h, and collected by centrifugation. The cell pellets were resuspended in phosphate-buffered saline (PBS) containing 1 μg ml^{-1} aprotinin, 1 μg ml^{-1} leupeptin, and 0.5 mM EDTA, and then sonicated. Insoluble material was removed by centrifugation. GST-fusion proteins were purified from the soluble extracts in a glutathione-Sepharose 4B column (Amersham Pharmacia Biotech), according to the manufacturer's instructions.

Chromatin immunoprecipitation (ChIP) analysis. ChIP assays were performed as described[5]. Immunoprecipitation of the crosslinked chromatin–protein complexes was done with the anti-Myc (9E10). PCR analyses of the 5′-flanking region of *Per2* gene were done using a primer set described previously[30], and SYBR Green I-based real-time quantitative PCR analyses were done as described[5].

30. Yoo, S. H. *et al.* A noncanonical E-box enhancer drives mouse Period2 circadian oscillations *in vivo*. *Proc. Natl Acad. Sci. USA* **102**, 2608–2613 (2005).

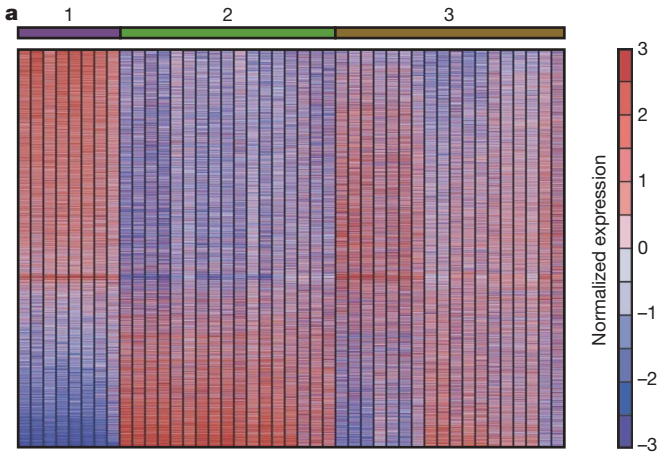
LETTERS

Distinct physiological states of *Plasmodium falciparum* in malaria-infected patients

J. P. Daily[1,3], D. Scanfeld[4], N. Pochet[4,5], K. Le Roch[6], D. Plouffe[7], M. Kamal[4], O. Sarr[8], S. Mboup[8], O. Ndir[9], D. Wypij[2], K. Levasseur[1], E. Thomas[4], P. Tamayo[4], C. Dong[1], Y. Zhou[7], E. S. Lander[4,10,11], D. Ndiaye[9], D. Wirth[1], E. A. Winzeler[7,12], J. P. Mesirov[4]* & A. Regev[4,10]*

Infection with the malaria parasite *Plasmodium falciparum* leads to widely different clinical conditions in children, ranging from mild flu-like symptoms to coma and death[1]. Despite the immense medical implications, the genetic and molecular basis of this diversity remains largely unknown[2]. Studies of *in vitro* gene expression have found few transcriptional differences between different parasite strains[3]. Here we present a large study of *in vivo* expression profiles of parasites derived directly from blood samples from infected patients. The *in vivo* expression profiles define three distinct transcriptional states. The biological basis of these states can be interpreted by comparison with an extensive compendium of expression data in the yeast *Saccharomyces cerevisiae*. The three states *in vivo* closely resemble, first, active growth based on glycolytic metabolism, second, a starvation response accompanied by metabolism of alternative carbon sources, and third, an environmental stress response. The glycolytic state is highly similar to the known profile of the ring stage *in vitro*, but the other states have not been observed *in vitro*. The results reveal a previously unknown physiological diversity in the *in vivo* biology of the malaria parasite, in particular evidence for a functional mitochondrion in the asexual-stage parasite, and indicate *in vivo* and *in vitro* studies to determine how this variation may affect disease manifestations and treatment.

To study the molecular basis of disease variation in malaria after infection with *P. falciparum*, we analysed the expression profiles of parasites derived directly from venous blood samples[4,5] of 43 patients residing in Senegal, with a diverse age range (8.3 ± 6.9 years (mean ± s.d.)), and illness severity (parasitaemia 5.5% ± 6.2%, haematocrit 32.3 ± 6.8 (means ± s.d.)). Although previous studies found little variation between expression profiles of different *P. falciparum* strains *in vitro*[3], we proposed that variation in the human



b

Variable	Cluster 1 (*n* = 8)	Cluster 2 (*n* = 17)	Cluster 3 (*n* = 18)
Age (years)	5 (2–15)	6 (3–6)	7 (4–15)
Parasitaemia (%)	2 (1–8)	4 (2–6)	3 (2–9)
Glucose (mg dl^{-1})	108 (94–113)	99 (88–121)	101 (54–13)
Haematocrit	39 (30–45)	33 (29–36)	30 (27–35)
Days ill	3 (3–3)	3 (2–3.8)	4 (3–6)*
Temperature (°C)	38 (37–40)	38 (37–39)	39 (38–40)*
IL-6 (pg ml^{-1})	53 (11–108)	29 (19–180)	190 (53–1,051)
IL-10 (pg ml^{-1})	561 (157–1,853)	216 (113–2,354)	4270 (1,349–8,088)*
TNF-α (pg ml^{-1})	69 (35–119)	52 (25–82)	67 (51–95)
TGF-α (pg ml^{-1})	29 (19–60)	33 (22–43)	87 (50–383)*
Tissue factor (pg ml^{-1})	169 (147–231)	181 (151–200)	345 (248–539)*
VCAM-1 (μg ml^{-1})	6 (3–8)	5 (3–7)	10 (6–18)*
Lymphotactin (pg ml^{-1})	197 (189–204)	200 (181–236)	265 (219–336)*

Figure 1 | *P. falciparum* expression profiles *in vivo*. **a**, NMF clustering of expression profiles. The expression values for 3,937 *P. falciparum* genes (rows) across 43 samples (columns) are shown. Genes with very low expression were thresholded to a minimum value and filtered to exclude those that showed little variation across samples (Methods). Samples were first clustered by NMF and the genes were then sorted by their discrimination between cluster 1 versus all other samples. Each gene's expression is normalized by mean centring and scaling (colour bar). The clustering identified three transcriptional states, two of which (clusters 1 and 2) are diametrically opposed and may represent a transcriptional shift. The number of clusters was determined objectively by the method, which does not force a structure on the data. The NMF clustering was repeated with samples derived from 2005 only ($n = 31$), and the cluster groups were unchanged (Supplementary Fig. 7). **b**, Clinical correlates of patients in each cluster. Shown are the median values and interquartile ranges of host demographic and selected laboratory values including cytokine measurements in the patients in each cluster. Statistically significant values (Mann–Whitney test with cluster 2 data as the reference group, $P < 0.05$) are designated by an asterisk. Cluster 3 is associated with significantly elevated inflammation markers, including duration of illness and body temperature and elevated levels of IL-6, IL-10, transforming growth factor (TGF)-α, tissue factor, vascular cell adhesion molecule (VCAM)-1 and lymphotactin. TNF, tumour necrosis factor.

[1]Department of Immunology and Infectious Disease, [2]Department of Biostatistics, Harvard School of Public Health, 665 Huntington Avenue, Boston, Massachusetts 02115, USA. [3]Department of Medicine, Brigham and Women's Hospital, 75 Francis Street, Boston, Massachusetts 02115, USA. [4]Broad Institute of Massachusetts Institute of Technology and Harvard University, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [5]FAS Center for Systems Biology, Harvard University, 7 Divinity Avenue, Cambridge, Massachusetts 02138, USA. [6]Department of Cell Biology and Neuroscience, 900 University Avenue, University of California, Riverside, California 92521, USA. [7]Genomics Institute of the Novartis Research Foundation, San Diego, California 92121, USA. [8]Laboratory of Bacteriology and Virology, [9]Department of Parasitology and Mycology, Dantec Hospital, Cheikh Anta Diop University, Dakar, BP 5005, Senegal. [10]Department of Biology, Massachusetts Institute of Technology, 31 Ames Street, Cambridge, Massachusetts 02139, USA. [11]The Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, Massachusetts 02142, USA. [12]Department of Cell Biology, The Scripps Research Institute, 10550 Torrey Pines Road, La Jolla, California 92037, USA.
*These authors contributed equally to this work.

host environment might affect *P. falciparum* biology and be reflected in its transcriptional profile.

We clustered the samples' expression profiles, using a non-negative matrix factorization (NMF) algorithm[6] (Fig. 1a, Supplementary Fig. 1 and Methods) and discovered that expression profiles cluster into three distinct groups. The profiles of samples in cluster 2 were similar to early ring-stage profiles of the 3D7 strain grown *in vitro*[7–9] (for example, Spearman rank correlation 0.54 on average compared with ref. 7; Supplementary Fig. 2 and Supplementary Note 1). Ring stages predominate in the peripheral blood, and these were the only stages we observed in blood smears from the 43 samples (Supplementary Fig. 3). In contrast, expression profiles of samples in clusters 1 and 3 were not similar to those of early rings (0.12 and 0.26) or late stages (0.06 and 0.01) of the asexual parasite life cycle *in vitro*, and were only weakly similar to profiles of other developmental states such as gametocytes[9] (0.31 and 0.23) or sporozoites (0.35 and 0.33; Supplementary Fig. 2 and Supplementary Note 1). They therefore represent novel transcriptional states. Profiles in clusters 1 and 2 are internally homogeneous and diametrically opposed, possibly reflecting a global transcriptional shift. Cluster 3 represents a third, distinct, pattern, although with more heterogeneity. Computational analysis indicates that profiles in cluster 3 are not a mixture of populations in cluster 1 and cluster 2 states (Supplementary Note 2).

The distinction between clusters 1 and 2 is not a reflection of patients' measured parameters, of parasite genotypes or of different life cycle stages. There were no statistically significant differences between the clusters with respect to patients' parameters, parasitological characterization (Fig. 1b), demographics or laboratory profiles. Parasite genotypes that identify distinct clones and number of clones in a single patient (MSP1/2) and chloroquine resistance (PFCRT K76T) showed no association with the clusters (data not shown). Furthermore, clusters 1 and 2 did not correlate with dates of sample collection, RNA isolation or oligonucleotide array hybridization. Examination of blood smears of each sample confirmed that only early ring stages were present (Supplementary Fig. 3) and the same clustering was observed with a set of 1,190 genes that do not vary during the parasite's asexual life cycle[7] (Supplementary Fig. 4).

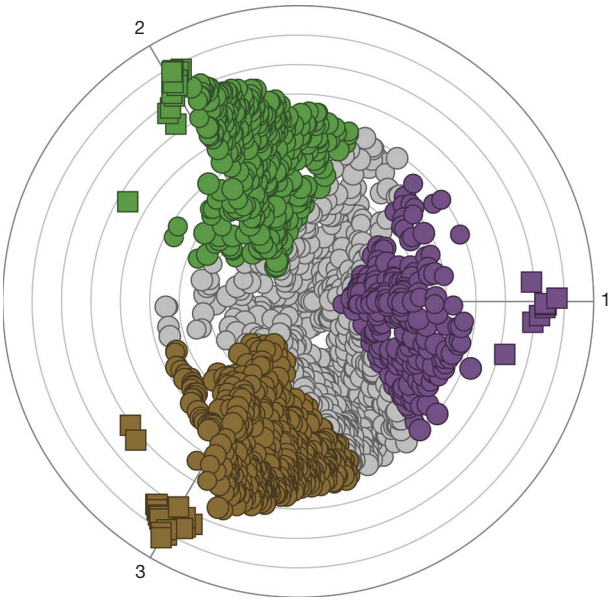


Figure 2 | Physiological characterization of *Plasmodium* profiles by cross-species projection. Shown is a radial plot mapping of 1,439 array experiments from *S. cerevisiae* (circles) projected onto the expression space defined by the three *P. falciparum* NMF clusters (purple, green and brown squares corresponding to *P. falciparum* samples from each of clusters 1, 2 and 3, respectively). Yeast experiments associated with each cluster (Brier score ≥ 0.4) are highlighted with the corresponding colour (Methods).

To identify the physiological basis of the distinct transcriptional states, we compared the *P. falciparum* expression patterns with a compendium of 1,439 published expression profiles from the yeast *S. cerevisiae* (Methods and Supplementary Table 1). We mapped 1,247 *S. cerevisiae* genes to their *P. falciparum* orthologues (Methods) and then scored each *S. cerevisiae* profile for its similarity to the three expression clusters (Methods). For each cluster in *P. falciparum*, we identified a set of similar *S. cerevisiae* profiles and examined their biological annotations. We also used Gene Set Enrichment Analysis (GSEA)[10] to test for the induction or repression of known pathways or functions (755 sets from *P. falciparum*; 328 sets from *S. cerevisiae*).

Each of the *P. falciparum* clusters was associated with a distinct set of *S. cerevisiae* responses (Fig. 2). Cluster 2 matched *S. cerevisiae* profiles associated with normal fermentative (glycolytic) growth

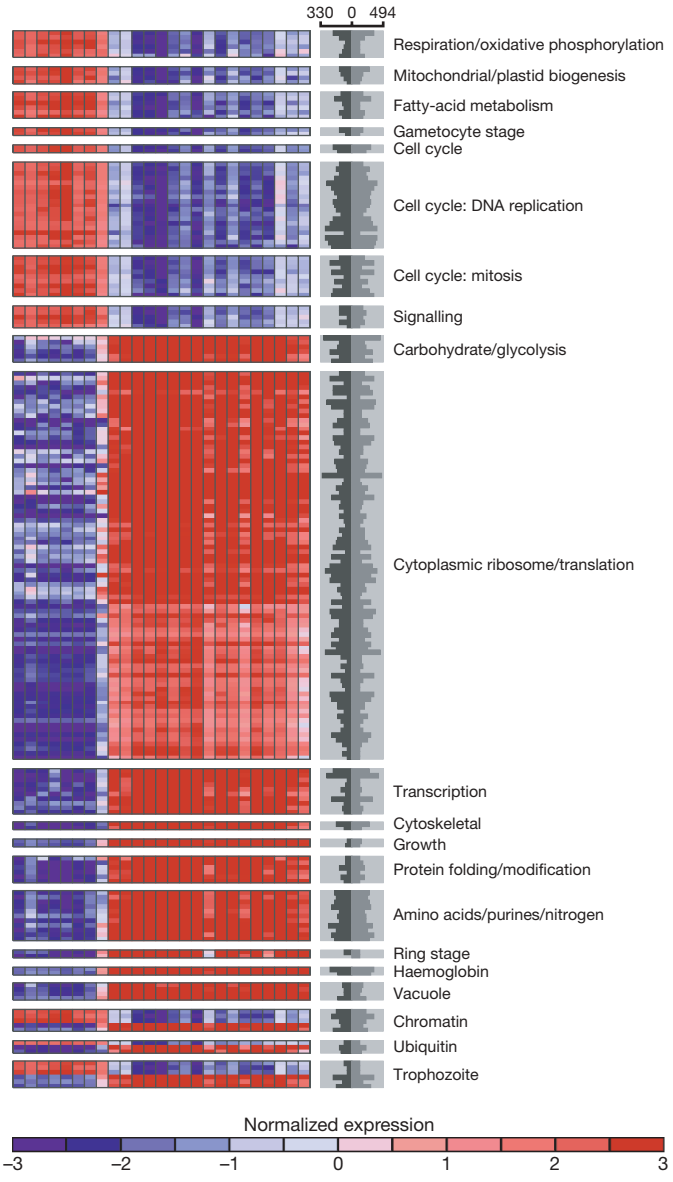


Figure 3 | Gene-set enrichment analysis of *P. falciparum* clusters. All the gene sets (rows) that differed significantly between cluster 1 and cluster 2 are shown, labelled by general categories. For each gene set, the mean expression of the 'leading-edge' genes (which supported the differential expression signature) in each experiment from the two clusters is shown (columns). The experiments are ordered as in Fig. 1. General biological categories describing the gene sets appear on the right; only gene sets with clear biological descriptions are included. Coloured bars indicate the number of genes in each gene set and in the leading edge.

(168/287 experiments, $P = 2.3 \times 10^{-23}$), cluster 1 matched profiles associated with starvation responses of *S. cerevisiae* (44/113, $P = 1.5 \times 10^{-7}$) as well as mutations in the general transcription machinery (23/53 experiments, $P = 2.8 \times 10^{-5}$). Cluster 3 was strongly associated with experiments on environmental stress in *S. cerevisiae* (278/438, $P = 4.6 \times 10^{-22}$).

This interpretation was also strongly supported by the induction of specific pathways and genes (Figs 3–5, Supplementary Table 2 and Supplementary Table 3). Cluster 2 showed induction of gene sets associated with glycolysis, amino-acid and nitrogen metabolism, and general growth processes such as nuclear transcription and cytoplasmic translation. By contrast, cluster 1 showed induction of gene sets associated with oxidative phosphorylation, respiration, mitochondrial biogenesis, the apicoplast, fatty-acid metabolism and genes involved in the uptake and metabolism of glycerol[11–13] (Figs 3–5, Supplementary Table 2 and Supplementary Fig. 5). Thus, parasites in cluster 1 may rely on alternative pathways of energy production through the use of substrates such as glycerol, lactic acid, other carbon sources or lipids present in the patient's blood. In addition, cluster 1 shows induction of genes related to invasion; this observation may be of clinical significance.

Cluster 1 shows induction of cell-cycle related modules of both DNA replication and mitotic functions (Fig. 3), although the parasites in these samples were in the early ring stage (Supplementary Fig. 3). This induction explains some of the weak similarity of cluster 1 to some profiles from later stages of the asexual life cycle[7,8] and from the sexual life cycle[9] (Supplementary Fig. 2 and Supplementary Note 1). However, cluster 1 does not directly correspond to these developmental stages. This can be readily seen by examining key processes that are coherently induced in cluster 1. Although particular subsets of genes within these processes are induced at various points in the asexual cycle, there is no stage in the cycle that shows coherent induction of the genes within each process or of the overall collection of processes (Supplementary Note 1, Supplementary Fig. 6 and Supplementary Table 8).

What is the biological basis for the difference between clusters 1 and 2? Parasites of the reference strain are typically grown *in vitro* under glucose-rich and microaerophilic conditions, and they depend on anaerobic glycolysis for energy[14]. It has been widely assumed that exclusive reliance on anaerobic glycolysis represents the physiology of the asexual parasite *in vivo*. Cluster 2 is consistent with such glycolytic growth *in vivo*.

In contrast, cluster 1 indicates that a starvation response can lead to a metabolic shift in the asexual stage of *P. falciparum* and that respiration and metabolism of alternative carbon sources may be important in parasite physiology *in vivo*. This suggests that the metabolism of *P. falciparum* is consistent with that of the *P. yoelii* and *P. berghei* model systems[15], which show active respiratory chains. Thus, parasites *in vivo* may exist in different states, as a result of varied oxygen or substrate levels. Although overall oxygen and substrate levels are tightly regulated in the human host, parasites are sequestered for half of their life cycle in the microvasculature, and oxygenation and substrate levels in this microenvironment can vary[16,17]. Furthermore, humans exhibit specific transcriptional changes when infected with *Plasmodium*[18]; our data indicate that the host environment may in turn affect parasite transcription.

Cluster 3 was strongly associated with *S. cerevisiae* profiles measured under environmental stress (for example heat shock, oxidative stress or osmotic stress) and also showed a clear correlation with the patients' clinical phenotypes. In particular (Fig. 1b), the patients have

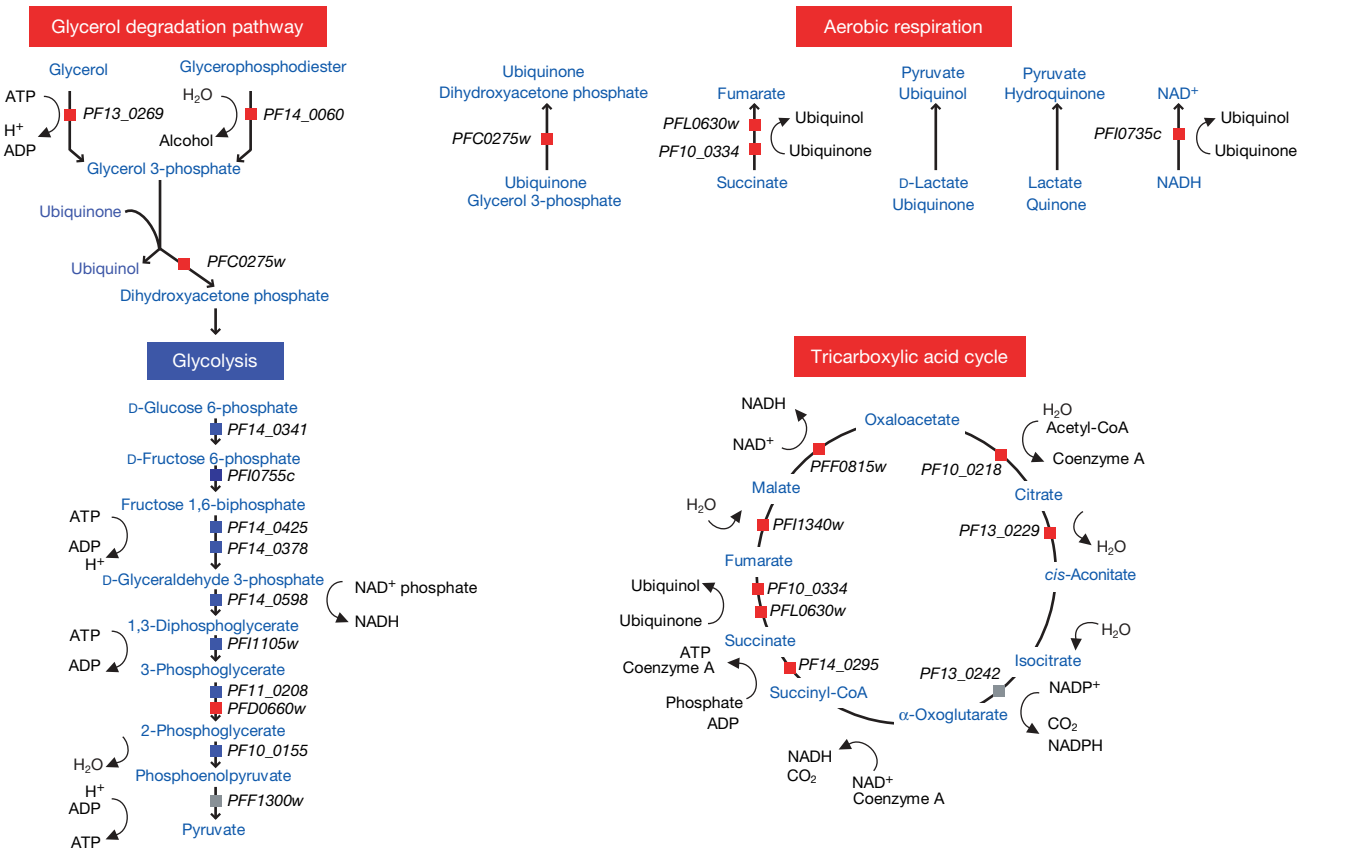


Figure 4 | Induction of respiratory metabolism and repression of glycolysis in cluster 1 versus cluster 2. Metabolic pathways derived from PlasmoCyc for glycolysis (glycolysis I), tricarboxylic acid cycle, aerobic respiration (electron donors reaction list) and glycerol degradation (glycerol degradation I) are shown[13]. The mean expression level for genes encoding the enzymes catalysing each reaction was calculated for cluster 1 and cluster 2. A ratio of expression for these values is indicated by colour bars. Red (blue) bars represent genes with at least twofold higher (lower) expression in cluster 1 versus cluster 2. Grey represents no change.

a higher temperature, greater inflammation and elevated levels of the cytokines interleukin (IL)-6 and IL-10, which have been associated with more severe outcomes[19]. It has previously been demonstrated that parasite biology can change in response to environmental cues[20]. Additional samples from patients with severe disease will be needed to understand the clinical significance of this cluster.

Epigenetic mechanisms may have a role in the establishment of these transcriptional shifts. First, cluster 1 profiles resemble those observed in *S. cerevisiae* single-gene knockouts in general transcription factors (for example subunits of the Mediator, TFIID and SAGA complexes). These may be critical for the establishment of distinct transcriptional programmes. Second, the transcript encoding the CCAAT-binding protein is significantly induced in cluster 1. This protein is orthologous to the key regulator of oxidative phosphorylation genes from yeast to humans[21,22]. This factor may have a similar role in *P. falciparum*. More broadly, we found marked differences between clusters 1 and 2 in the expression of multiple genes encoding histones and chromatin modifiers (Supplementary Table 4), which may be critical for the establishment of stable and distinct transcriptional programmes in *P. falciparum*. Reproducing this transcriptional shift *in vitro* is critical for discovering its physiological and mechanistic basis.

Our observations about the apparent starvation response in samples in cluster 1 raise possible connections with gametogenesis. First, starvation responses typically cause yeast and other eukaryotic microbes to finish asexual growth and undergo meiosis. Second, respiratory and mitochondrial functions are known to be induced in gametocytes that have multiple mitochondria and higher oxygen consumption[23]. Third, the expression profiles in cluster 1 are more similar to late stages of *in vitro* gametogenesis[9] than those in the other clusters, although the similarity is weak. Fourth, the expression of known gametogenesis genes[9] is higher in cluster 1 samples than in cluster 2 (data not shown). Malaria parasites in the ring state choose between sexual and asexual fates long before morphological differences are apparent. Because gametocytes are isolated by the indiscriminate killing of immature sexual and asexual parasites, we know little about the metabolism or transcriptional programmes of these early sexual stages. It will be interesting to investigate whether the starvation response in cluster 1 may lead to a shift *in vivo* to a sexual form that allows the parasite to escape its starved host by transmitting through the mosquito vector into a new host. This hypothesis could be tested through studies of starvation *in vitro* and of parasite stages *in vivo*.

Pathogenesis studies in other systems have shown that organisms have distinct biology *in vivo* in comparison with *in vitro* models, and that some of these differences relate to virulence[24]. Little is known about the biology of *Plasmodium* residing in the human circulation.

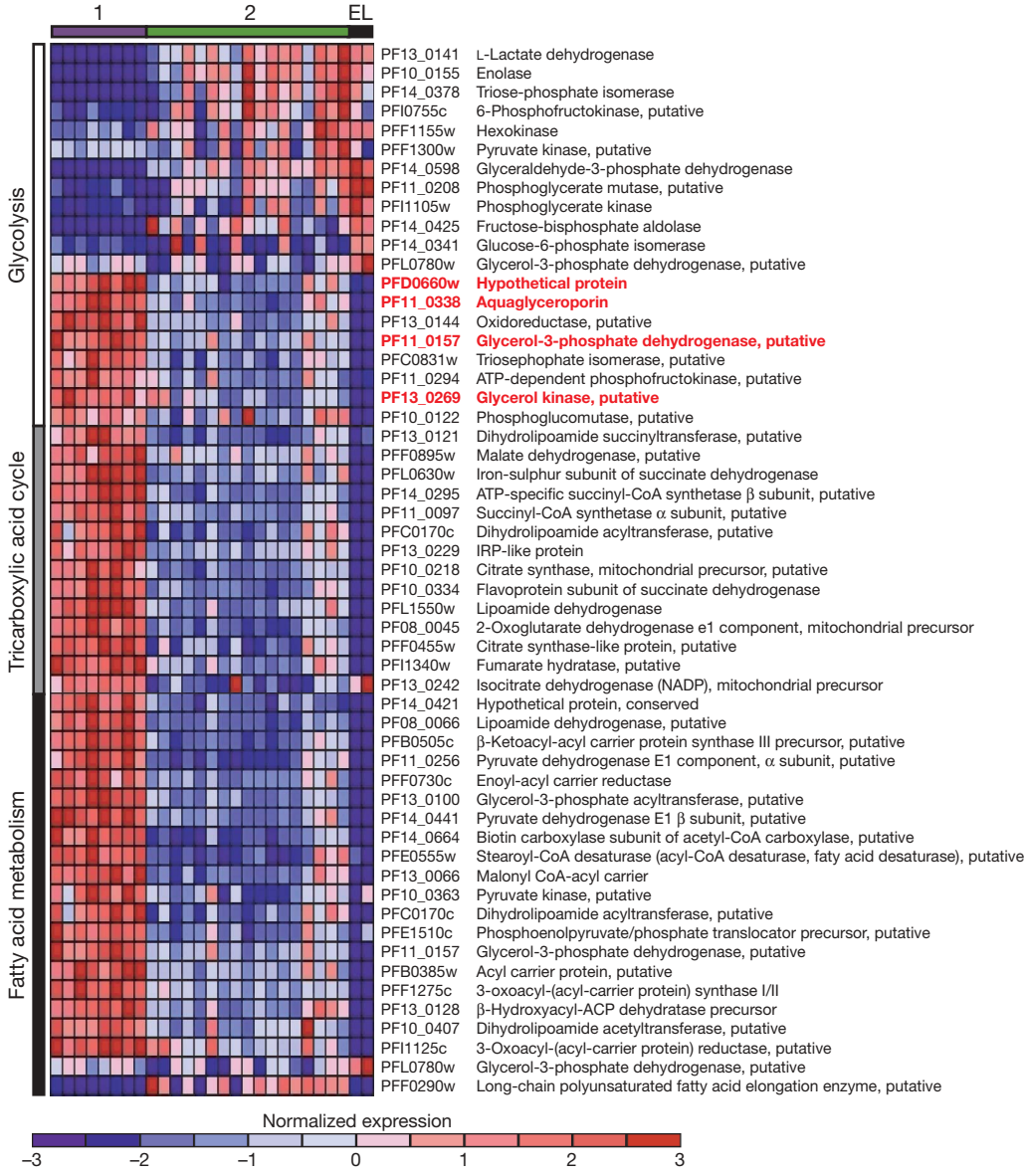


Figure 5 | Expression of glycolysis, tricarboxylic acid cycle and fatty acid metabolism genes in clusters 1 and 2. Relative expression of genes participating in major metabolic pathways. Hierarchical clustering of the expression values of the genes participating in glycolysis, the tricarboxylic acid cycle and fatty-acid metabolism[30] in samples in cluster 1, cluster 2 and 3D7 early (E) and late (L) ring stages[7]. Names of glycolysis genes important for glycerol metabolism, including those encoding a glycerol transporter (PF11_0338) and aerobic glycerol catabolism enzymes (PF11_0660w, PF11_0157 and PF13_0269) are shown in red. The mean expression values for each gene in each cluster are reported in Supplementary Table 7. The relatively high expression level of genes involved in glycerol degradation and fatty-acid metabolism in cluster 1 compared with their expression in cluster 2 may suggest the use of alternative carbon sources for energy production.

Our results show that the *Plasmodium* parasite exists in the human host in at least three distinct physiological states, apparently related to glycolytic growth, a starvation response and a general (non-nutritional) stress response. The relationships between these states and the course of clinical disease remain to be elucidated. Nevertheless, it is notable that cluster 1 shows strong induction of genes encoding proteins involved in invasion pathways, and cluster 3 is significantly associated with host inflammation. These novel states may result in enhanced virulence and the generation of metabolites such as reactive oxygen species, or in the consumption of substrates that could affect the host and contribute to disease severity[17]. Finally, if the distinct profiles represent persistent physiological differences, they may identify novel drug targets for malaria or may indicate possible alternative therapies.

METHODS SUMMARY

Patient population and sample handling. Venous blood samples from *P. falciparum*-infected patients in Senegal were directly added to Tri-Reagent BD (Molecular Research Center). This cohort consisted of patients who presented to the district hospital in Velingara, Senegal, with fever and symptoms suggestive of malaria. Enrolment criteria consisted of a *P. falciparum* infection of at least 1% of red blood cells. RNA was isolated, and steady-state parasite messenger RNA levels in 43 samples were determined with a custom-made Affymetrix chip based on the 3D7 genome as reported previously[7].

Transcriptional analysis. The patient-derived transcriptional profiles were normalized with each other and with previously published *in vitro* data sets[7–9] to allow direct comparisons. Samples were clustered by using NMF[6], which finds a small number of gene combinations (metagenes) that best capture the behaviour of an expression data set. The number of clusters was determined using consensus clustering and maximizing the cophenetic correlation coefficient. Gene sets that are differentially expressed between clusters were identified by GSEA[10], on the basis of a weighted Kolmogorov–Smirnov-like statistic. To project yeast expression data onto our parasite data set we first identified 1,247 *S. cerevisiae* genes that have *P. falciparum* orthologues. We then used metagene projection[25] combined with a Support Vector Machine predictor to project 1,439 previously published[26] *S. cerevisiae* expression profiles into the three metagene factor NMF representations described above (Supplementary Table 1) with a confidence level determined by a Brier score[25]. Experiments scoring highly in a given factor were associated with the *P. falciparum* cluster represented by that factor. We then used a hypergeometric enrichment test to identify biological conditions enriched in the profiles associated with each cluster. The complete data, gene sets, and associated analyses are available from http://carrier.gnf.org/publications/PatientProfiling and http://www.broad.mit.edu/compbio/pub/malaria

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 17 May; accepted 26 September 2007.
Published online 28 November 2007.

1. White, N. in *Manson's Tropical Diseases* 21st edn (eds Cook, G. C. & Zumla, A. I.) 1205–1295 (Elsevier Science and W. B. Saunders, Edinburgh, 2002).
2. Greenwood, B., Marsh, K. & Snow, R. Why do some African children develop severe malaria? *Parasitol. Today* **7,** 277–281 (1991).
3. Llinas, M., Bozdech, Z., Wong, E. D., Adai, A. T. & DeRisi, J. L. Comparative whole genome transcriptome analysis of three *Plasmodium falciparum* strains. *Nucleic Acids Res.* **34,** 1166–1173 (2006).
4. Daily, J. P. *et al. In vivo* transcriptional profiling of *P. falciparum*. *Malar. J.* **3,** 30 (2004).
5. Daily, J. P. *et al. In vivo* transcriptome of *Plasmodium falciparum* reveals overexpression of transcripts that encode surface proteins. *J. Infect. Dis.* **191,** 1196–1203 (2005).
6. Brunet, J. P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl Acad. Sci. USA* **101,** 4164–4169 (2004).
7. Le Roch, K. G. *et al.* Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* **301,** 1503–1508 (2003).
8. Bozdech, Z. *et al.* The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biol.* **1,** E5 (2003).
9. Young, J. A. *et al.* The *Plasmodium falciparum* sexual development transcriptome: a microarray analysis using ontology-based pattern identification. *Mol. Biochem. Parasitol.* **143,** 67–79 (2005).
10. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102,** 15545–15550 (2005).
11. Hansen, M., Kun, J. F., Schultz, J. E. & Beitz, E. A single, bi-functional aquaglyceroporin in blood-stage *Plasmodium falciparum* malaria parasites. *J. Biol. Chem.* **277,** 4874–4882 (2002).
12. Promeneur, D. *et al.* Aquaglyceroporin PbAQP during intraerythrocytic development of the malaria parasite *Plasmodium berghei*. *Proc. Natl Acad. Sci. USA* **104,** 2211–2216 (2007).
13. Yeh, I., Hanekamp, T., Tsoka, S., Karp, P. D. & Altman, R. B. Computational analysis of *Plasmodium falciparum* metabolism: organizing genomic information to facilitate drug discovery. *Genome Res.* **14,** 917–924 (2004).
14. Lang-Unnasch, N. & Murphy, A. D. Metabolic changes of the malaria parasite during the transition from the human to the mosquito host. *Annu. Rev. Microbiol.* **52,** 561–590 (1998).
15. Uyemura, S. A., Luo, S., Vieira, M., Moreno, S. N. & Docampo, R. Oxidative phosphorylation and rotenone-insensitive malate- and NADH-quinone oxidoreductases in *Plasmodium yoelii yoelii* mitochondria *in situ*. *J. Biol. Chem.* **279,** 385–393 (2004).
16. Tsai, A. G., Johnson, P. C. & Intaglietta, M. Oxygen gradients in the microcirculation. *Physiol. Rev.* **83,** 933–963 (2003).
17. Planche, T. & Krishna, S. Severe malaria: metabolic complications. *Curr. Mol. Med.* **6,** 141–153 (2006).
18. Ockenhouse, C. F. *et al.* Common and divergent immune response signaling pathways discovered in peripheral blood mononuclear cell gene expression patterns in presymptomatic and clinically apparent malaria. *Infect. Immun.* **74,** 5561–5573 (2006).
19. Lyke, K. E. *et al.* Serum levels of the proinflammatory cytokines interleukin-1β (IL-1β), IL-6, IL-8, IL-10, tumor necrosis factor α, and IL-12(p70) in Malian children with severe *Plasmodium falciparum* malaria and matched uncomplicated malaria or healthy controls. *Infect. Immun.* **72,** 5630–5637 (2004).
20. Udomsangpetch, R. *et al.* Febrile temperatures induce cytoadherence of ring-stage *Plasmodium falciparum*-infected erythrocytes. *Proc. Natl Acad. Sci. USA* **99,** 11825–11829 (2002).
21. Olesen, J., Hahn, S. & Guarente, L. Yeast HAP2 and HAP3 activators both bind to the CYC1 upstream activation site, UAS2, in an interdependent manner. *Cell* **51,** 953–961 (1987).
22. Becker, D. M., Fikes, J. D. & Guarente, L. A cDNA encoding a human CCAAT-binding protein cloned by functional complementation in yeast. *Proc. Natl Acad. Sci. USA* **88,** 1968–1972 (1991).
23. Krungkrai, J., Prapunwattana, P. & Krungkrai, S. R. Ultrastructure and function of mitochondria in gametocytic stage of *Plasmodium falciparum*. *Parasite* **7,** 19–26 (2000).
24. Mahan, M. J., Slauch, J. M. & Mekalanos, J. J. Selection of bacterial virulence genes that are specifically induced in host tissues. *Science* **259,** 686–688 (1993).
25. Tamayo, P. *et al.* Metagene projection for cross-platform, cross-species characterization of global transcriptional states. *Proc. Natl Acad. Sci. USA* **104,** 5959–5964 (2007).
26. Marion, R. M. *et al.* Sfp1 is a stress- and nutrient-sensitive regulator of ribosomal protein gene expression. *Proc. Natl Acad. Sci. USA* **101,** 14315–14322 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the villagers and health care workers in Velingara, Senegal, for their participation in and support of this project; T. Taylor for critical advice and encouragement; G. Chechik, N. Barkai and O. Rando for providing parts of the compiled expression compendium for *S. cerevisiae*; and J. Bistline for assistance with the figures. J.P.D. is supported by the National Institute of Allergy and Infectious Diseases. N.P. is a Henri Benedictus Fellow of the King Baudouin Foundation and the Belgian American Educational Foundation. P.T. is supported by the National Institutes of Health (NIH). D.F.W. is supported by the Ellison Medical Foundation, the NIH, the Exxon Mobil Foundation and the Harvard School of Public Health. E.A.W. is supported by the Keck Foundation, the Novartis Research Foundation and the NIH. J.P.M. and D.S. are supported by the NIH and the National Science Foundation. A.R. is supported by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund. The field work was supported by the Fogarty International from the NIH, the NIH Malaria Diversity grant, the Exxon Mobil Foundation, the Ellison Medical Foundation, the Burroughs–Welcome Fund and the Broad Institute of MIT and Harvard.

Author Contributions. D.S., N.P. and K.L.R. contributed equally to this work.

Author Information The array data are deposited in the Gene Expression Omnibus under accession number GSE9152. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.P.M. (mesirov@broad.mit.edu) or A.R. (aregev@broad.mit.edu).

METHODS

Patient population and study site. A field site was established in Velingara, a hyperendemic village in eastern Senegal, with peak transmission from October to December and an entomological inoculation rate of over 100 (ref. 27). Samples were collected during two transmission seasons, October to November in 2004 and 2005. Patients who required hospitalization or who appeared severely ill were enrolled in 2004. This cohort included two patients with asymptomatic hypoglycaemia, one patient with respiratory acidosis and one patient with coma. To obtain a larger sample size, all patients fulfilling enrolment criteria were enrolled in 2005, including those with minimal symptoms. Patients who presented to the hospital in Velingara were triaged by the local nurse to undergo malaria smear if they had symptoms suggestive of malaria. Enrolment criteria consisted of a *P. falciparum* infection without a second species noted on thin smear of 1% parasitaemia or greater. Of 1,187 patients screened for *P. falciparum* infection, 412 had a positive blood smear for *P. falciparum* and 95 fulfilled the enrolment criteria; and all consented to the study. After informed consent had been obtained, patients underwent venipuncture and one or two blood tubes (10–20 ml) coated with K_3EDTA was collected. Tri-reagent BD (Molecular Research Center) was added within 5–10 min after blood collection. All samples were processed by a single person. The mixture was maintained at 4 °C until each evening, when it was placed in liquid nitrogen. Haematocrit was measured by microhaematocrit centrifugation. The remaining sample was centrifuged and divided into aliquots for serum studies, parasite cryopreservation, short-term culture, and application to filter paper for later DNA extraction. Cytokines, soluble endothelial-cell ligands and markers of inflammation were analysed from patient serum with a multiplex sandwich ELISA (Searchlight). Serum glucose levels were determined in Boston (on an Olympus AU 2700 analyser) from the transported frozen serum aliquots. Protocols were approved by the Harvard School of Public Health Human Subjects Committee and Senegal Ministry of Health Research Ethics Committee.

Detection of mRNA transcripts. The samples were shipped to Boston in liquid nitrogen, thawed at room temperature and total RNA was isolated in accordance with the manufacturer's protocol (Tri reagent BD). Twelve samples from 2004 and 31 samples from 2005 that demonstrated sharp ribosomal bands on a denaturing agarose gel stained with ethidium bromide were selected for hybridization. Steady-state parasite mRNA levels were determined with a custom-made Affymetrix chip based on the 3D7 genome as reported previously[7]. Hybridizations were performed on three separate dates.

Data filtering and normalization. Each transcript was assigned a relative expression unit (EU) using MOID, used as reported previously[28]. A filtered gene list containing 3,937 genes was generated by thresholding gene expression levels to a minimum of 50 EU and removing any genes that varied less than threefold or 100 EU across the data set. To minimize potential effects of different dates of collection and hybridizations, the data was rank ordered by expression level and each gene was given an ordinal value. The published 3D7 reference strain data were processed in the same manner to allow comparisons[7,9].

NMF clustering. The 43 *P. falciparum*-derived expression profiles were clustered by using NMF as described previously[6] using the GenePattern software[29]. We chose a three-cluster solution, yielding the three distinct groups of samples, based on cluster-membership stability using consensus clustering (Supplementary Fig. 1). To determine whether the clustering is robust to the date of sample collection, we repeated NMF clustering using only the 31 samples collected in 2005 (Supplementary Fig. 7); these yielded the same results. The matrices derived from the NMF factorization give a description of the data in terms of three metagenes (three positive linear combinations of all the genes). Using the previously described metagene projection methodology[25], we created an NMF projection of the data into three metagene factors, each corresponding to a compact representation of the associated cluster. To improve this projection, we equalized the number of samples to eight in each cluster. Cluster 1 had only eight samples. Because of the high degree of heterogeneity in cluster 3, we chose the eight samples in the other clusters to represent the widest range of behaviour. We then recomputed the NMF projection and used this final map to project the *S. cerevisiae* profiles into the same three-metagene representation.

Identification of *S. cerevisiae*–*P. falciparum* orthologues. We used the Kyoto Encyclopedia of Genes and Genomes[30] SSDB database (KEGG) to find reciprocal best pairs of *P. falciparum*–*S. cerevisiae* genes with Smith–Waterman similarity scores of 100 or more. In all, 24% of the *P. falciparum* genome and 21% of the *S. cerevisiae* genome were included in the matched pairs.

Gene sets. *P. falciparum* gene annotations and pathways were obtained from KEGG, PlasmoDB, Hagai Ginsburg's Malaria Metabolic Pathways[31] and Gene Ontology (GO)[30,32,33]. Yeast gene modules were constructed by following the procedure of ref. 34 using a yeast expression compendium described below and a total of 3,395 gene classes, including 1,794 from the GO[35] hierarchy, 87 from KEGG, 107 from the BioCyc database[36], 1,022 from the MIPS database of manually curated protein complexes[37], 310 from a data set describing the genes whose promoters are bound by various transcription factors, 70 from a data set describing the genes that harbour a given *cis*-regulatory element in their promoter[38], and 5 from a data set describing the genes whose RNA is bound by the RNA-binding proteins from the PUF family[39]. The yeast modules were mapped onto *P. falciparum* genes on the basis of the orthology relations described above.

GSEA. GSEA was performed as described previously[10]. In brief, the procedure assesses whether an *a priori* defined set of genes shows statistically significant, concordant differences between two biological states. Given a data set and two classes, genes are ranked on the basis of the correlation between their expression and the distinction between the two classes. GSEA uses a weighted Kolmogorov–Smirnov-like statistic to calculate an enrichment score that reflects the degree to which a gene set is overrepresented at the extremes of the entire ranked list. Within a gene set there is a leading-edge subset, which is defined as the genes that appear before the point in which the running sum enrichment score reaches its maximum deviation from zero. Because of the small number of samples in our study we estimated significance on the basis of a gene label (rather than class label) permutation. In our analyses we considered gene sets with a nominal *P* value below 0.01 and a false discovery rate (FDR) below 0.01 to be significant. The FDR for the invasion gene set is 0.07. We tested a total of 755 gene sets defined in *P. falciparum* and 328 sets originally defined in *S. cerevisiae*.

***S. cerevisiae* expression compendium.** A compendium of 1,439 previously published *S. cerevisiae* expression profiles was compiled from the literature as described previously[26] (Supplementary Table 1). Each experiment was manually annotated according to the experimental conditions, based on 20 categories (Supplementary Table 5); in addition each experiment was automatically annotated on the basis of the coherent induction or repression of each *S. cerevisiae* gene set (above) by following the procedure of ref. 34 (Supplementary Table 6).

Projection of *S. cerevisiae* experiments. Each *S. cerevisiae* expression profile was mapped into the *P. falciparum* gene space on the basis of the orthology assignments. Next, a Support Vector Machine predictor was used to project the *S. cerevisiae* expression profiles into the three-metagene-factor NMF representation described above. Experiments scoring highly in a given factor could be related to the *P. falciparum* cluster represented by that factor. Using a modified Brier skill score[25], we measured a confidence level for each of these predictions. For each factor (cluster) we defined an associated set of the *S. cerevisiae* experiments that scored 0.4 or more for that factor. Next, we tested which array annotations were significantly enriched in each set of *S. cerevisiae* arrays by using the hypergeometric distribution to calculate a *P* value. The reported results were robust to the particular NMF model that we employed and to the threshold of Brier score used (ranging from 0.25 to 0.75).

Molecular analysis. The number of clones was determined by assessing MSP-1 and MSP-2 allelic variants as described previously[40]. Determination of the chloroquine-resistance-associated *pfcrt* K76T mutation was performed with PCR and restriction-fragment-length polymorphism using 3D7 chloroquine-sensitive and W2 chloroquine-resistant genomic DNA as controls[41]. To determine whether there were statistical differences in host features between clusters, the Mann–Whitney test was performed with Stata (version 9.0).

27. Faye, O. *et al.* Comparison of the transmission of malaria in 2 epidemiological patterns in Senegal: the Sahel border and the Sudan-type savanna. *Dakar Med.* **40,** 201–207 (1995).
28. Zhou, Y. & Abagyan, R. Match-Only Integral Distribution (MOID) algorithm for high-density oligonucleotide array analysis. *BMC Bioinformatics* **3,** 3 (2002).
29. Reich, M. *et al.* GenePattern 2.0. *Nature Genet.* **38,** 500–501 (2006).
30. Kanehisa, M. A database for post-genome analysis. *Trends Genet.* **13,** 375–376 (1997).
31. Ginsburg, H. Progress in *in silico* functional genomics: the Malaria Metabolic Pathways database. *Trends Parasitol.* **22,** 238–240 (2006).
32. Bahl, A. *et al.* PlasmoDB: the *Plasmodium* genome resource. A database integrating experimental and computational data. *Nucleic Acids Res.* **31,** 212–215 (2003).
33. Zhou, Y. *et al. In silico* gene function prediction using ontology-based pattern identification. *Bioinformatics* **21,** 1237–1245 (2005).
34. Segal, E., Friedman, N., Koller, D. & Regev, A. A module map showing conditional activity of expression modules in cancer. *Nature Genet.* **36,** 1090–1098 (2004).
35. Ashburner, M., Mungall, C. J. & Lewis, S. E. Ontologies for biologists: a community model for the annotation of genomic data. *Cold Spring Harb. Symp. Quant. Biol.* **68,** 227–235 (2003).
36. Karp, P. D. *et al.* Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.* **33,** 6083–6089 (2005).
37. Mewes, H. W. *et al.* MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res.* **34,** D169–D172 (2006).
38. Harbison, C. T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431,** 99–104 (2004).

39. Gerber, A. P., Herschlag, D. & Brown, P. O. Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. *PLoS Biol.* **2,** E79 (2004).
40. Snounou, G. *et al.* Biased distribution of *msp1* and *msp2* allelic variants in *Plasmodium falciparum* populations in Thailand. *Trans. R. Soc. Trop. Med. Hyg.* **93,** 369–374 (1999).
41. Happi, C. T. *et al.* Molecular analysis of *Plasmodium falciparum* recrudescent malaria infections in children treated with chloroquine in Nigeria. *Am. J. Trop. Med. Hyg.* **70,** 20–26 (2004).

Vol 450|13 December 2007|doi:10.1038/nature05992

LETTERS

A viral microRNA functions as an orthologue of cellular miR-155

Eva Gottwein[1], Neelanjan Mukherjee[2], Christoph Sachse[4], Corina Frenzel[4], William H. Majoros[5], Jen-Tsan A. Chi[1,5], Ravi Braich[7], Muthiah Manoharan[7], Jürgen Soutschek[7], Uwe Ohler[3,5,6] & Bryan R. Cullen[1]

All metazoan eukaryotes express microRNAs (miRNAs), roughly 22-nucleotide regulatory RNAs that can repress the expression of messenger RNAs bearing complementary sequences[1]. Several DNA viruses also express miRNAs in infected cells, suggesting a role in viral replication and pathogenesis[2]. Although specific viral miRNAs have been shown to autoregulate viral mRNAs[3,4] or downregulate cellular mRNAs[5,6], the function of most viral miRNAs remains unknown. Here we report that the miR-K12-11 miRNA encoded by Kaposi's-sarcoma-associated herpes virus (KSHV) shows significant homology to cellular miR-155, including the entire miRNA 'seed' region[7]. Using a range of assays, we show that expression of physiological levels of miR-K12-11 or miR-155 results in the downregulation of an extensive set of common mRNA targets, including genes with known roles in cell growth regulation. Our findings indicate that viral miR-K12-11 functions as an orthologue of cellular miR-155 and probably evolved to exploit a pre-existing gene regulatory pathway in B cells. Moreover, the known aetiological role of miR-155 in B-cell transformation[8–10] suggests that miR-K12-11 may contribute to the induction of KSHV-positive B-cell tumours in infected patients.

Inspection of mature KSHV miR-K12-11 and cellular miR-155 reveals significant homology, including the entire seed region that is often critical for mRNA target recognition[7]; that is, nucleotides 2–8 (Fig. 1a). miR-155, the product of the *bic* gene[10], is overexpressed in several types of B-cell lymphoma, and its transgenic expression in mice causes B-cell malignancies[9]. miR-155 expression is induced in activated B cells, T cells and macrophages[11–13], and miR-155 knockout mice have impaired immune functions[14,15]. Given the emerging importance of miR-155 in cancer and B-cell function, we asked whether miR-K12-11 functions as an orthologue of miR-155.

We first prepared transductants of the KSHV-negative human B-cell line BJAB expressing physiological levels of miR-K12-11. A miR-K12-11 expression cassette was placed 3′ to the *AcGFP* (a variant of green fluorescent protein cloned from *Aequorea coerelescens*) open reading frame (ORF) present in the lentiviral vector pNL-SIN-CMV-AcGFP (Fig. 1b). Cells transduced with this vector express transcripts that function as *AcGFP* mRNAs and as primary miRNAs (pri-miRNAs)[16]. Transduced BJAB cells were sorted to generate pools expressing only AcGFP or expressing AcGFP and miR-K12-11. Expression of miR-K12-11 was confirmed by primer extension (Fig. 1c, lanes 1–8, and Supplementary Fig. 2a, b) and by knockdown of an indicator bearing perfectly complementary sites[17] (Supplementary Fig. 2c–e). The level of expression and activity of miR-K12-11 in transduced BJAB cells was comparable to that observed in the B-cell line BC-1 (Fig. 1c, lane 15, and Supplementary Fig. 2), which is latently infected with KSHV[18].

Cytoplasmic RNA was isolated from BJAB cells and analysed on microarrays in three independent experiments. Expression of

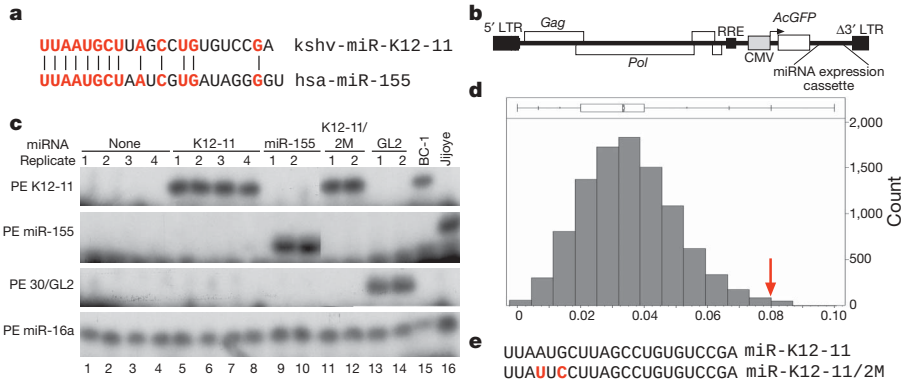


Figure 1 | Expression of miR-K12-11 and miR-155 in BJAB cells. **a**, Alignment of mature miR-K12-11 and miR-155. **b**, Schematic of the lentiviral miRNA expression vector[16]. LTR, long terminal repeat. **c**, Primer extension (PE) analysis of miRNA expression in BJAB transductants and control cell lines BC-1 and Jijoye for miR-K12-11, miR-155, miR-30/GL2 and cellular miR-16. **d**, Null distribution of miR-K12-11/miR-155 seed abundance in 10,000 randomly sampled sets of 150 3′ UTRs, each from the space of our analysis. The arrow marks the abundance of heptamer seed matches in the 3′ UTRs of the top-scoring 150 downregulated genes in miR-K12-11 transductants, indicating a significant ($P \leq 0.0057$) enrichment of potential targets for miR-K12-11/miR-155. Analysis for hexamer seed matches also identified a significant enrichment ($P \leq 0.001$) (not shown). **e**, Mutations introduced to create miR-K12-11/2M are highlighted.

[1]Department of Molecular Genetics and Microbiology, [2]University Program in Genetics and Genomics, [3]Department of Biostatistics & Bioinformatics, Duke University Medical Center, Durham, North Carolina 27710, USA. [4]Cenix BioScience GmbH, Tatzberg 47, 01307 Dresden, Germany. [5]Institute for Genome Sciences and Policy, [6]Department of Computer Science, Duke University, Durham, North Carolina 27708, USA. [7]Alnylam Pharmaceuticals, Inc., 300 3rd Street, Cambridge, Massachusetts 02142, USA.

miR-K12-11 resulted in the downregulation of 181 mRNAs and the upregulation of 86 mRNAs ($|T\text{-score}| \geq 2.86$, $P \leq 0.01$). mRNAs bearing at least one seed match to miR-K12-11 were significantly enriched in the top-scoring 150 mRNAs downregulated in miR-K12-11-expressing cells, in comparison with randomly sampled 3′ untranslated regions (UTRs) from all mRNAs in our data set (Fig. 1d). As a result of stringent criteria for 3′ UTR selection, this data set probably underestimates the number of seed matches to miR-K12-11 in cellular 3′ UTRs. We therefore used the Entrez database to manually check mRNAs that were significantly downregulated for additional seed matches. Downregulated mRNAs carrying at least a six-nucleotide seed match to miR-K12-11 were considered candidate direct targets and are listed in Supplementary Table 1. Using Gene Set Enrichment Analysis[19], we observed that computer-predicted targets for cellular miR-155 were significantly enriched in mRNAs downregulated by miR-K12-11 (false discovery rate ≤ 0.003; Supplementary Table 2)[20]. The downregulation of many of these mRNAs by miR-K12-11 was not statistically significant, on the basis of our *T*-score cutoff, suggesting low or inconsistent levels of downregulation by miR-K12-11. However, inhibition of gene expression by miRNAs can occur entirely at the translational level, with little or no reduction in mRNA abundance[21,22].

We next evaluated whether candidate mRNA targets for miR-K12-11 were also downregulated by miR-155 using quantitative RT–PCR analysis of RNA samples prepared from duplicate sets of BJAB transductants (Fig. 1c). As a specificity control, we also analysed a miRNA, miR-K12-11/2M, that bears two mutations in the miR-K12-11 seed region (Fig. 1e). Because the pri-miRNA stem was adjusted accordingly (Supplementary Fig. 1), the expression level of this mutant was comparable to that of wild-type miR-K12-11 (Fig. 1c, lanes 11 and 12). miR-155 was expressed from a 300-base-pair (bp) fragment of *bic* exon 2 (ref. 10) and its expression level in BJAB (Fig. 1c, lanes 9 and 10) matched that observed in the transformed B-cell line Jijoye (Fig. 1c, lane 16). BJAB cells normally express undetectable levels of endogenous miR-155 (Fig. 1c). As a further control, we also included BJAB transductants expressing an artificial miRNA targeting luciferase (miR-30/GL2).

Because the set of potential mRNA targets of miR-K12-11 was diverse and did not reveal an obvious function, candidate mRNAs were picked on the basis of *T*-score, known functions, and/or the number and quality of matches to miR-K12-11. We also included several predicted mRNA targets for miR-155 that did not seem to be significantly downregulated by miR-K12-11. Supplementary Table 3 gives the known functions of these candidate targets, as well as their full names and the extent of seed pairing to miR-K12-11 and miR-155. As shown in Fig. 2 and Supplementary Fig. 3, quantitative reverse transcriptase-mediated polymerase chain reaction (qRT–PCR) analysis revealed that almost all the transcripts analysed were indeed downregulated by miR-K12-11 but not by miR-K12-11/2M, thereby proving that downregulation depends on an intact seed sequence. Importantly, expression of miR-155 downregulated these mRNAs to the same degree as did miR-K12-11, thus strongly suggesting that miR-K12-11 does indeed phenocopy miR-155 (Fig. 2).

We next asked whether the downregulation of target genes by miR-K12-11 and miR-155 was direct by cloning 12 candidate 3′ UTR sequences 3′ to the firefly luciferase ORF (*Fluc*). The resulting constructs and the unmodified vector were co-transfected into 293T cells together with a *Renilla* luciferase internal control and pNL-SIN-CMV-AcGFP constructs expressing no miRNA, miR-K12-11, miR-K12-11/2M or miR-155. In all 12 cases, Fluc expression from constructs bearing candidate 3′ UTR sequences was downregulated equivalently by miR-K12-11 and miR-155 (Fig. 3), whereas miR-K12-11/2M had no effect. Consistent with the idea that target gene repression also occurs by inhibition of translation, there was no good correlation between the extent of downregulation observed by qRT–PCR (Fig. 2) and the indicator assay (Fig. 3). For example, one predicted target for miR-155, the transcription factor BTB and CNC homology 1 (BACH1), was only modestly downregulated at the mRNA level, but repression of the indicator bearing the *BACH1* 3′ UTR was about 85% for both miR-155 and miR-K12-11.

To show that equivalent downregulation by miR-K12-11 and miR-155 also occurs with endogenous proteins, we performed western blot analyses for two targets, BACH1 and Fos. Whole-cell extracts of duplicate BJAB transductants (Fig. 1c) were probed with antibodies specific for BACH1. Because miR-155 is upregulated in stimulated macrophages[13], we also generated transductants of the monocytic cell line THP-1 expressing each miRNA and confirmed expression by using primer extension (Supplementary Fig. 4). BACH1 protein expression was about fivefold lower in the BJAB transductants expressing either miR-K12-11 or miR-155 when compared with control transductants (Fig. 4a), whereas expression in THP-1 cells was inhibited about threefold (Fig. 4b). Fos expression was analysed after serum starvation of BJAB transductants followed by induction with 12-*O*-tetradecanoylphorbol-13-acetate (TPA).

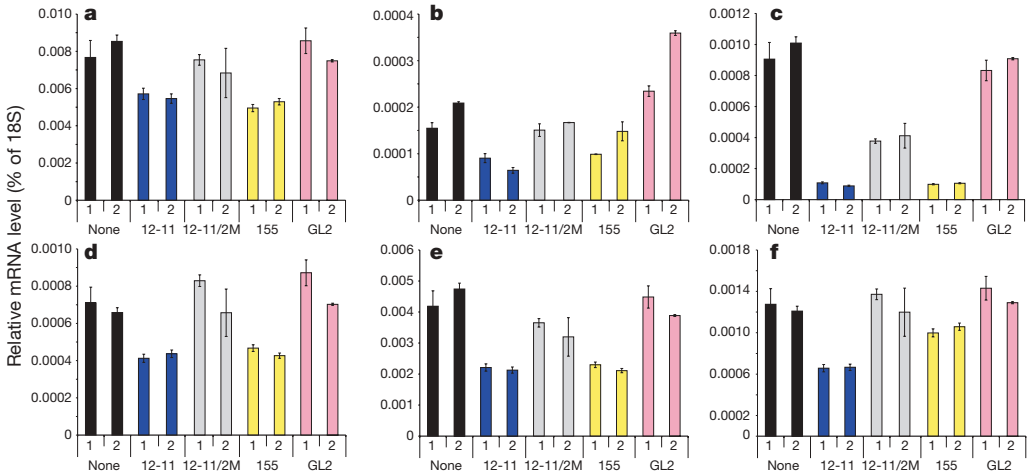


Figure 2 | miR-K12-11 and miR-155 regulate a common set of mRNAs. Real-time qRT–PCR analysis of total RNA derived from two independent replicates of BJAB transductants, expressing no miRNA, miR-K12-11, miR-K12-11/2M, miR-155 or miR-30/GL2, for six candidate mRNA targets of miR-K12-11 and miR155. Relative RNA abundance is shown as a percentage of the level of 18S ribosomal RNA, and error bars (s.d.) are derived from quadruplicate 18S rRNA replicates. mRNAs tested included BACH1 (**a**), Fos (**b**), BIRC4BP (**c**), AGTRAP (angiotensin II receptor-associated protein) (**d**), SAMHD1 (SAM domain and HD domain 1) (**e**) and RFK (riboflavin kinase)(**f**). Results for further candidate targets are shown in Supplementary Fig. 3.

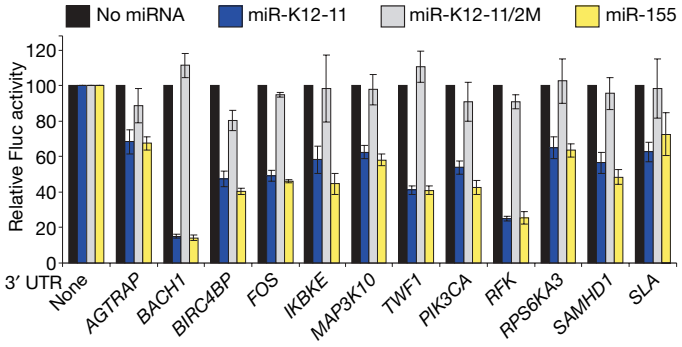


Figure 3 | Direct and equivalent regulation of candidate 3′ UTRs by miR-K12-11 and miR-155. Indicator vectors carrying no additional sequences, or candidate 3′ UTR sequences inserted 3′ to *Fluc*, were co-transfected with an internal RLuc control vector and either empty pNL-SIN-CMV-AcGFP (black bars), pNL-SIN-CMV-AcGFP expressing miR-K12-11 (blue), miR-K12-11/2M (grey) or miR-155 (yellow). Dual luciferase assays were performed 48 h later. For each pNL-SIN-CMV-AcGFP construct, *Fluc* to *Rluc* ratios were first normalized to the value obtained for the empty indicator vector and then to the activities obtained with pNL-SIN-CMV-AcGFP, which were set at 100%. Error bars (s.d.) are derived from three independent experiments. BACH1, IKBKE, MAP3K10 (mitogen-activated protein kinase kinase kinase 10), SLA and RPS6KA3 (ribosomal protein S6 kinase) are predicted targets for miR-155 (Supplementary Table 2).

Again, levels of Fos protein expression after induction were consistently about 2.5-fold lower in transductants expressing either miR-K12-11 or miR-155 (Fig. 4c).

If miRNAs are important in viral replication *in vivo*, then agents that specifically block their function might represent novel antiviral agents. We treated latently KSHV-infected BCBL-1 cells with an antagomir[23] antisense to miR-K12-11 and asked whether this would increase the expression of Fos and/or BACH1 protein. As shown in Fig. 4d, this antagomir specifically enhanced Fos expression about 2.6-fold while selectively inhibiting miR-K12-11 expression and function (Supplementary Fig. 5). In contrast, BACH1 expression was only modestly enhanced (data not shown). Further analysis revealed that the *BACH1* 3′ UTR confers downregulation not only by miR-K12-11 but also by miR-K12-1 and miR-K12-6 (Supplementary Fig. 6). Therefore, although antagomirs can inhibit viral miRNA function, their effectiveness may be compromised by the overlapping activities of viral miRNAs.

Previous analyses have shown that certain viral miRNAs downregulate the expression of mRNA targets transcribed from the opposite strand of the viral DNA genome[3,4] or inhibit the expression of cellular mRNAs by binding to novel 3′ UTR targets that are not used by cellular miRNAs[5,6]. Here we provide evidence for a third model, namely that viral miRNAs can function as orthologues of cellular miRNAs and thereby downregulate the expression of numerous cellular mRNAs by means of target sites that are generally evolutionarily conserved.

The evidence presented here argues strongly that miR-K12-11 functions as an orthologue of cellular miR-155 (Figs 2–4). This similarity undoubtedly reflects the identical seed region present in miR-K12-11 and miR-155 (Fig. 1a). Nevertheless, the fact that the non-seed regions of miR-155 and miR-K12-11 are different does raise the possibility that they might not share all mRNA targets[24]. Indeed, although almost all mRNAs analysed responded identically to miR-155 and miR-K12-11 (see, for example, Figs 2 and 3), we did occasionally notice slight differences in the degree of silencing (for example, mRNA encoding Src-like-adaptor (SLA); Supplementary Fig. 4). Nevertheless, taken together, our data do demonstrate that miR-155 and miR-K12-11 regulate an analogous set of mRNAs and agree with earlier reports documenting the critical role of seed pairing in mRNA target selection[7].

miR-155/bic overexpression is observed in many human B-cell lymphomas[10] and induces B-cell malignancies in mice and chickens[8,9]. miR-155 is induced on T-cell or B-cell activation[11,12], and miR-155 knockout mice have impaired immune function[14,15]. The set of mRNA targets identified here includes targets with known roles in B-cell function (for example SLA) and innate immunity (PIK3CA (phosphoinositide-3-kinase, catalytic α subunit), IKBKE (IκB kinase ε) and Fos), pro-apoptotic (BIRC4BP/XAF1 (XIAP associated factor-1)) and cell-cycle-regulatory (Fos) functions, as well as transcription factors (BACH1, Fos and HIVEP2 (human immunodeficiency virus type I enhancer binding protein 2)). At this point, it is unclear which of the many genes regulated by miR-155 and by miR-K12-11 provide a replicative advantage to KSHV. Given the apparent role of miR-155 in the development of B-cell tumours, miR-K12-11 expression in latently KSHV-infected B cells may contribute to the increased incidence of B-cell tumours seen in KSHV-infected patients[25]. Although the distantly related γ-herpes virus Epstein–Barr virus (EBV) also expresses viral miRNAs[2,3], these show no homology to miR-155. However, a recent report[26] showing that EBV activates endogenous miR-155 expression in infected B cells suggests that EBV may have evolved an alternative strategy to achieve the same end result.

Analysis of currently known viral miRNAs reveals that the seed homology observed with miR-K12-11 and miR-155 is not unique. Supplementary Fig. 7 lists several viral miRNAs that display seed

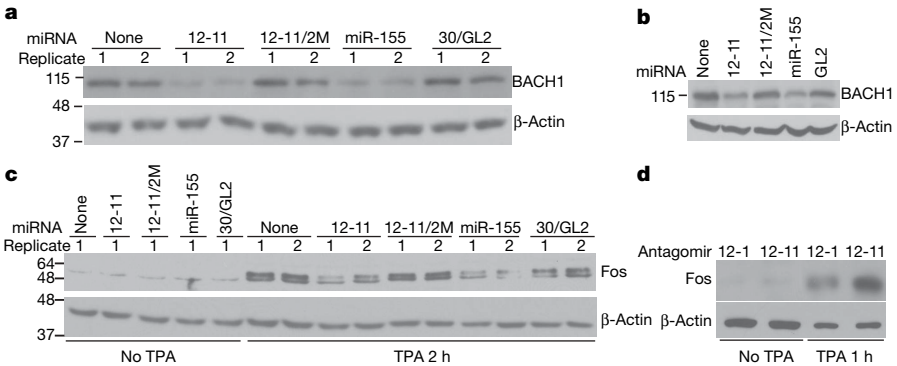


Figure 4 | Endogenous BACH1 and Fos proteins are repressed by both miR-K12-11 and miR-155. **a**, Western blot analysis of BACH1 protein expression in independently generated BJAB transductants expressing no miRNA, miR-K12-11, miR-K12-11/2M, miR-155 or miR-30/GL2. **b**, Western blot analysis of BACH1 protein expression in THP-1 transductants. **c**, Western blot analysis of Fos protein. Replicate BJAB transductants expressing no miRNA, miR-K12-11, miR-K12-11/2M, miR-155 or miR-30/GL2 were serum starved for 26 h and then treated with serum-free RPMI medium or TPA for 2 h. **d**, Fos expression in KSHV-infected cells is rescued by a miR-K12-11-specific antagomir. BCBL-1 cells were serum starved for 25 h in the presence of 1 μM antagomir against miR-K12-1 or miR-K12-11 and then incubated for 1 h in serum-free medium in the presence or absence of 20 ng ml^{-1} TPA. Numbers at the left of panels are the molecular masses of marker proteins in kDa.

homology either to cellular miRNAs or to miRNAs encoded by distantly related viruses. Although the significance of these homologies remains unknown, these observations do raise the possibility that the similar activity noted for cellular miR-155 and for viral miR-K12-11 is only the first example of a more general phenomenon.

METHODS SUMMARY

pNL-SIN-CMV-AcGFP-based lentiviral miRNA expression vectors were generated as described[16], except that expression cassettes were placed 3′ to the *AcGFP* ORF. miR-K12-11 and miR-30/GL2 were expressed from artificial miR-30-based expression cassettes[22,27]. miR-155 was expressed from a roughly 300-bp segment of the *bic* gene[10]. BJAB cells were infected and sorted 48 h after infection. At 12–16 days after transduction, gene expression analysis of ten independent BJAB pools, expressing AcGFP only or AcGPF and miR-K12-11, was performed with Human Operon v3.0.2 arrays. Normalization was performed with arrayMagic[28] and data analysis was conducted with GenePattern[29]. 3′ UTR analysis for miR-K12-11/miR-155 seed matches was performed with an in-house pipeline[30]. Gene Set Enrichment Analysis[19] was used to test whether specific gene sets were enriched in the set of downregulated genes. Candidate 3′ UTR sequences were cloned 3′ to an SV40 early promoter-driven *Fluc* ORF. Indicator assays were conducted in 293T cells co-transfected with *Fluc* vectors carrying candidate 3′ UTRs or the parental vector, an *RLuc*-based internal control vector, and pNL-SIN-CMV-AcGFP-based miRNA expression vectors. Dual luciferase assays were performed 48 h after transfection[17]. Antagomirs were synthesized as described[23] and delivered into BCBL-1 cells by incubation in serum-free medium for 25 h.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 27 August; accepted 10 October 2007.

1. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116,** 281–297 (2004).
2. Cullen, B. R. Viruses and microRNAs. *Nature Genet.* **38** (Suppl.), S25–S30 (2006).
3. Pfeffer, S. *et al.* Identification of virus-encoded microRNAs. *Science* **304,** 734–736 (2004).
4. Sullivan, C. S., Grundhoff, A. T., Tevethia, S., Pipas, J. M. & Ganem, D. SV40-encoded microRNAs regulate viral gene expression and reduce susceptibility to cytotoxic T cells. *Nature* **435,** 682–686 (2005).
5. Stern-Ginossar, N. *et al.* Host immune system gene targeting by a viral miRNA. *Science* **317,** 376–381 (2007).
6. Samols, M. A. *et al.* Identification of cellular genes targeted by KSHV-encoded microRNAs. *PLoS Pathog.* **3,** e65 (2007).
7. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120,** 15–20 (2005).
8. Clurman, B. E. & Hayward, W. S. Multiple proto-oncogene activations in avian leukosis virus-induced lymphomas: evidence for stage-specific events. *Mol. Cell. Biol.* **9,** 2657–2664 (1989).
9. Costinean, S. *et al.* Pre-B cell proliferation and lymphoblastic leukemia/high-grade lymphoma in Eμ-miR155 transgenic mice. *Proc. Natl Acad. Sci. USA* **103,** 7024–7029 (2006).
10. Eis, P. S. *et al.* Accumulation of miR-155 and BIC RNA in human B cell lymphomas. *Proc. Natl Acad. Sci. USA* **102,** 3627–3632 (2005).
11. van den Berg, A. *et al.* High expression of B-cell receptor inducible gene *BIC* in all subtypes of Hodgkin lymphoma. *Genes Chromosom. Cancer* **37,** 20–28 (2003).
12. Haasch, D. *et al.* T cell activation induces a noncoding RNA transcript sensitive to inhibition by immunosuppressant drugs and encoded by the proto-oncogene, *BIC*. *Cell. Immunol.* **217,** 78–86 (2002).
13. O'Connell, R. M., Taganov, K. D., Boldin, M. P., Cheng, G. & Baltimore, D. MicroRNA-155 is induced during the macrophage inflammatory response. *Proc. Natl Acad. Sci. USA* **104,** 1604–1609 (2007).
14. Thai, T. H. *et al.* Regulation of the germinal center response by microRNA-155. *Science* **316,** 604–608 (2007).
15. Rodriguez, A. *et al.* Requirement of bic/microRNA-155 for normal immune function. *Science* **316,** 608–611 (2007).
16. Gottwein, E. & Cullen, B. R. Protocols for expression and functional analysis of viral microRNAs. *Methods Enzymol.* **427,** 229–243 (2007).
17. Gottwein, E., Cai, X. & Cullen, B. R. A novel assay for viral microRNA function identifies a single nucleotide polymorphism that affects Drosha processing. *J. Virol.* **80,** 5321–5326 (2006).
18. Cai, X. *et al.* Kaposi's sarcoma-associated herpesvirus expresses an array of viral microRNAs in latently infected cells. *Proc. Natl Acad. Sci. USA* **102,** 5570–5575 (2005).
19. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102,** 15545–15550 (2005).
20. Xie, X. *et al.* Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc. Natl Acad. Sci. USA* **104,** 7145–7150 (2007).
21. Olsen, P. H. & Ambros, V. The lin-4 regulatory RNA controls developmental timing in *Caenorhabditis elegans* by blocking LIN-14 protein synthesis after the initiation of translation. *Dev. Biol.* **216,** 671–680 (1999).
22. Zeng, Y., Wagner, E. J. & Cullen, B. R. Both natural and designed micro RNAs can inhibit the expression of cognate mRNAs when expressed in human cells. *Mol. Cell* **9,** 1327–1333 (2002).
23. Krutzfeldt, J. *et al.* Silencing of microRNAs *in vivo* with 'antagomirs'. *Nature* **438,** 685–689 (2005).
24. Grimson, A. *et al.* MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell* **27,** 91–105 (2007).
25. Cesarman, E. & Knowles, D. M. Kaposi's sarcoma-associated herpesvirus: a lymphotropic human herpesvirus associated with Kaposi's sarcoma, primary effusion lymphoma, and multicentric Castleman's disease. *Semin. Diagn. Pathol.* **14,** 54–66 (1997).
26. Jiang, J., Lee, E. J. & Schmittgen, T. D. Increased expression of microRNA-155 in Epstein–Barr virus transformed lymphoblastoid cell lines. *Genes Chromosom. Cancer* **45,** 103–106 (2006).
27. Silva, J. M. *et al.* Second-generation shRNA libraries covering the mouse and human genomes. *Nature Genet.* **37,** 1281–1288 (2005).
28. Buness, A., Huber, W., Steiner, K., Sultmann, H. & Poustka, A. arrayMagic: two-colour cDNA microarray quality control and preprocessing. *Bioinformatics* **21,** 554–556 (2005).
29. Reich, M. *et al.* GenePattern 2.0. *Nature Genet.* **38,** 500–501 (2006).
30. Majoros, W. H. & Ohler, U. Spatial preferences of microRNA targets in 3′ untranslated regions. *BMC Genomics* **8,** 152 (2007).
31. Lee, M. T., Coburn, G. A., McClure, M. O. & Cullen, B. R. Inhibition of human immunodeficiency virus type 1 replication in primary macrophages by using Tat- or CCR5-specific small interfering RNAs expressed from a lentivirus vector. *J. Virol.* **77,** 11964–11972 (2003).
32. Rubinson, D. A. *et al.* A lentivirus-based system to functionally silence genes in primary mammalian cells, stem cells and transgenic mice by RNA interference. *Nature Genet.* **33,** 401–406 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank T. Curran for rabbit anti-Fos antiserum. Flow cytometry was performed by J. Whitesides in the Duke Center for AIDS Research BSL3 Flow Cytometry Core Facility. Microarrays were performed in the Duke Microarray Facility. This research was supported by a National Institutes of Health grant to B.R.C. and by a Feodor Lynen Fellowship to E.G. U.O. is an Alfred P. Sloan Fellow in Computational Molecular Biology.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the paper on www.nature.com/nature. Correspondence and requests for materials should be addressed to B.R.C. (culle002@mc.duke.edu).

METHODS

MicroRNA expression cassettes. MicroRNA expression cassettes were placed into the 3′ UTR of the *AcGFP* gene in the context of the pcDNA3-based vector pcDNA3/AcGFP. The AcGFP coding region was PCR amplified with the primer pair 129/131. The resulting PCR product was digested with *Not*I, blunt-ended with Klenow enzyme and then cut with *Hin*dIII. pcDNA3 was cut with *Eco*RI, blunt-ended with Klenow enzyme, cut with *Hin*dIII and ligated to the PCR product.

Because miR-K12-11 expression cassettes derived from KSHV genomic sequences did not yield high levels of miR-K12-11 after transduction (data not shown), the sequence of mature KSHV miR-K12-11 was embedded into a fragment of the pri-miRNA gene for miR-30. This strategy has previously been validated and is now widely employed for the expression of small RNAs[22,27]. First, 5′- and 3′-flanking regions of miR-30 were amplified from BJAB genomic DNA by using primer pairs 76/77 (5′-flanking region) and 78/79 (3′- flanking region) and cut with *Mfe*I and *Xho*I or with *Xho*I and *Bam*HI, respectively. Both fragments were ligated into the *Eco*RI and *Bam*HI sites of a shuttle vector, pLNCX2M. pLNCX2M is a modified version of the MLV based retroviral vector pLNCX2 (BD Biosciences) and contains *Eco*RI and *Bam*HI sites upstream of a CMV-promoter-driven Neo cassette. This arrangement of the miR-30 flanking regions was originally described by Silva *et al.*[27] and results in unique *Xho*I and *Eco*RI (introduced with primer 78) sites that allow the insertion of miRNA precursor hairpin sequences. The precursor hairpins for miR-K12-11 and miR-30/GL2 were designed to express the miRNA from the 3′ arm of the pre-miRNA hairpin. The passenger strand was adjusted to contain one bulge (at nucleotide 9 from the 5′ end of the mature miRNA sequence). The stem-loop sequences used for expression of miR-K12-11 and miR-30/GL2 are shown in Supplementary Fig. 1. Two primers covering each miRNA stem-loop coding sequence (miR-K12-11, primer pair 117/118; miR-30/GL2, primer pair 119/120) were designed to be complementary over the central 31 nucleotides. Extension of the annealed primers with *Pfu* polymerase yielded the entire extended pre-miRNA stem-loop coding sequence. The resulting fragment was cut with *Xho*I and *Eco*RI and inserted between the miR-30 flanking regions. The same strategy was used to clone expression cassettes for miR-K12-2 (primers 103/104), miR-K12-4-3p (primers 107/108), miR-K12-5 (primers 109/110), miR-K12-6-3p (primers 125/126) and miR-K12-7 (primers 127/128).

miR-30-based miRNA expression cassettes were amplified from the pLNCX2M shuttle vectors described above using primers 140/141. A ~300-bp fragment of the *BIC* gene (corresponding to nucleotides 150–449 of the RNA; accession number AF402776) was cloned from BJAB cDNA with primers 485/486. Expression cassettes for miR-K12-1 (primers 198/199), miR-K12-3 (primers 200/201), miR-K12-8 (primers 204/205), miR-K12-9 (primers 206/207) and miR-K12-10 (primers 208/209) were PCR-amplified from the KSHV genome using BC-1 genomic DNA. 5′ primers contained an *Nru*I site and 3′ primers contained *Xba*I and *Eco*RV sites. PCR products were cut with *Nru*I and *Xba*I and inserted into the *Eco*RV and *Xba*I sites of pcDNA3/AcGFP, thus regenerating an *Eco*RV site upstream of the *Xba*I site to facilitate the insertion of further miRNA expression cassettes (see below).

The miR-K12-11/2M expression cassette was generated by using overlap PCR. Using the miR30/K12-11 expression cassette described above as template, two overlapping PCR products were generated with primer pairs 140/390R and 141/390F and used as templates in a second round of PCR with primer pair 140/141. The resulting expression cassette (miR30K11/2M) was inserted into pcDNA3/AcGFP as described above.

In THP-1 cells, tandem miRNA expression cassettes were used to allow higher expression levels of the miRNA at lower AcGFP expression levels. The miRNA in question was amplified from the corresponding pcDNA3/AcGFP construct using primers 140/141, cut with *Nru*I and *Xba*I and cloned into the *Eco*RV and *Xba*I sites of the appropriate pcDNA3/AcGFP vectors containing one miRNA expression cassette.

To clone pNL-SIN-CMV-AcGFP and derivatives, the pcDNA3/AcGFP vector and its derivatives containing one or more miRNA expression cassettes were cut with *Xba*I, blunt-ended with Klenow enzyme, and then cut with *Mlu*I, yielding a fragment containing part of the CMV promoter, the AcGFP coding region and the miRNA expression cassette(s). This fragment was used to replace the corresponding fragment of pNL-SIN-CMV-BLR[31], which was cut with *Xho*I, blunt-ended with Klenow enzyme and then cut with *Mlu*I.

pL/SV40 RL and FL indicator vectors. Indicator and control vectors to test candidate cellular 3′ UTRs were based on the self-inactivating lentiviral vector pLL3.7 (ref. 32). First, a polylinker was inserted between the *Apa*I and *Eco*RI sites of pLL3.7 (primers 238/39), introducing the unique restriction sites *Apa*I, *Bam*HI, *Xho*I, *Xba*I, *Not*I and *Eco*RI. The resulting vector (pL) retained only the 5′ promoter sequences and regulatory sequences, as well as the WRE element and the 3′ SIN long terminal repeat. Next, fragments containing the SV40 promoter and RLuc or FLuc (luc^+ gene) coding regions were inserted by using *Bam*HI and *Nhe*I (SV40 promoter) and *Nhe*I and *Xho*I restriction sites (*luc* ORFs). The SV40 promoter was amplified from pcDNA3 by PCR with primers 256 and 257. The RLuc fragment was obtained by digesting the previously described vector pNL-SIN-CMV-RLuc with *Nhe*I and *Xho*I. The luc^+ ORF was amplified from pGL3-CMV with PCR primers 363 and 364. The resulting vectors were named pL/SV40/Rluc and pL/SV40/GL3 and contain unique *Xho*I, *Xba*I, *Not*I and *Eco*RI sites downstream of the *luc* ORFs to facilitate 3′ UTR cloning. Throughout this study, pL/SV40/Rluc served as negative control vector and pL/SV30/GL3 was used to insert candidate 3′ UTRs. Candidate 3′ UTR sequences were amplified from BJAB cDNA or genomic DNA with the following primer pairs: 376/377 (IKBKE), 380/381 (PIK3CA), 400/401 (Fos), 425/428 (BIRC4BP), 447/448 (RFK), 449/450 (TWF1). 451/452 (SLA), 453/454 (SAMHD1), 457/458 (AGTRAP), 462/463 (RPS6KA3), 464/465 (MAP3K10) and 501/504 (BACH1). pNL-SIN-CMV-FL and pNL-SIN-CMVRL as well as the indicator vectors bearing two sites perfectly complementary to each KSHV miRNA were described previously[17].

Generation of BJAB transductants. Lentiviral vectors were produced by transfection of 293T cells as described[31] and used to transduce about 10^6 BJAB or THP-1 cells at a cell concentration of about 5×10^5 ml^{-1}. The next day, media were exchanged; 48 h after transduction, cells were collected by centrifugation and resuspended in RPMI medium containing 2 mM EDTA. AcGFP-expressing cells were sorted (using the 488-nm line of a 20-mW laser) and analysed with a BD FACSAria cell sorter with DiVa software (BD Biosciences). In experiment 1, 60,000 BJAB cells expressing only AcGFP or AcGFP and miR-K12-11 were sorted and, after sorting, were split into three biological replicates. In experiments 2 and 3, BJAB cells were transduced in three or four independent replicates, respectively, and ~40,000 cells were collected for each replicate. Cell populations of similar mean fluorescence intensities were collected for all samples. In experiment 3, cells were resorted on day 6 after transduction to ensure comparable AcGFP expression levels. Cytoplasmic RNA for microarray analysis was prepared with the RNeasy Mini kit (Qiagen) and harvested on days 12 (experiment 2) and 16 (experiments 1 and 3) after transduction of BJAB cells. Indicator assays shown in Supplementary Fig. 2 were conducted on the day of RNA preparation (experiment 3) or one day before RNA preparation (experiment 1 and 2).

For THP-1, cells were sorted three times to achieve similar levels of AcGFP expression levels.

RNA preparation and primer extension. For analysis of miRNA expression by primer extension, total RNA was prepared with TRIzol reagent as instructed, and 10 μg of RNA was used per reaction. Primer extension was performed with the Promega Primer Extension System. Probe sequences were as follows: for miR-16a, cgccaatatttacgtg; miR-K11 and miR-K11/2M, tcggacacaggctaag; miR-155, cccctatcacgattagc; miR30/GL2, tcacgtacgcggaata. Independently, miR-K12-11 expression from the miR30-based expression cassette described above was validated by northern analysis (not shown). Taken together, these results prove that the miRNA expressed from our miR30-based miR-K12-11 expression cassette is identical to miR-K12-11.

Spotted microarray, RNA and microarray probe preparation and hybridization. Arrays were printed at the Duke Microarray Facility with the Genomics Solutions OmniGrid 300 Arrayer. The arrays contain the Human Operon v3.0.2 arrays (Oligo Source) that possess 34,602 unique optimized 70-mers. RNA quality was ascertained with an Agilent 2100 bioanalyser (Agilent Technologies). Cytoplasmic RNA (10 μg) from each sample and the reference (Universal Human Reference RNA; Stratagene) were hybridized to oligo(dT) primers at 65 °C and then incubated at 42 °C for 2 h in the presence of reverse transcriptase, Cy5-dUTP or Cy3-dUTP and Cy5-dCTP or Cy3-dCTP, and a deoxynucleotide mix. In all cases, BJAB-derived RNA samples were labelled with Cy5 and reference samples were labelled with Cy3. NaOH was used to destroy residual RNA. Sample and reference cDNAs were then pooled, purified with QIAquick Purification Columns (Qiagen), mixed with hybridization buffer (50% formamide, 5 × SSC and 0.1% SDS), COT-1 DNA and polydeoxyadenylic acid to limit non-specific binding, and heated to 95 °C for 2 min. This mixture was pipetted onto a microarray slide and hybridized overnight at 42 °C on the MAUI hybridization system (BioMicro Systems). The array was then washed at increasing stringencies and scanned on a GenePix 4000B microarray scanner (Axon Instruments). All protocols are available in greater detail on the Duke Microarray Facility Website (http://microarray.genome.duke.edu/services/spotted-arrays/protocols). Array results were submitted to the GEO database.

Microarray normalization and analysis. All arrays were subjected to background subtraction followed by loess normalization within each array and scale normalization across all arrays with the arrayMagic package in R[28]. The KNN impute package in GenePattern[29] was used to impute missing data if a probe had intensity values for at least half the samples. Otherwise the probes were excluded

from analysis. Replicate probes were collapsed to one probe corresponding to the median value of all the replicates. Probe IDs that had a *t*-score of more than 4.032 (roughly $P \leq 0.01$) from a two-sided *t*-test, calculated with the Comparative Marker Selection package from GenePattern, comparing AcGFP-only-expressing cell populations with the unmodified BJAB cell line were considered to be activated by AcGFP and excluded from our further analysis (895 probes). At this stage, 23,330 probes remained and represent the set on which the analysis was conducted; *t*-scores were calculated for comparisons between cell populations expressing AcGFP only to those expressing AcGFP and miR-K12-11 by using the Comparative Marker Selection package from GenePattern.

miRNA target mapping. We applied an in-house computational pipeline[30] to analyse a high-quality set of human 3′ UTR sequences, based on the hg18 assembly, for the presence of sites complementary to miR-K12-11. The set of 3′ UTRs were first mapped from human Refseq IDs to Human Operon v3.0.2 probe IDs using the array annotation table downloaded from the Operon website. All successfully mapped 3′ UTRs were analysed for exact hexamer and heptamer matches to the reverse complement of the miRNA seed (positions 2–7 or 2–8 from the 5′ end of the mature miR-K12-11 sequence, respectively). Probes that contained at least one seed match in any transcript isoform were regarded as hits, without any requirements regarding conservation of seed matches in related species. These sets of UTRs and seed matches were used to build a null distribution and assess the significance of miR-K12-11 target-site frequency in the top list of downregulated genes.

Gene Set Enrichment Analysis and enrichment of K12-11 seed matches. To identify sets of genes that showed a correlated change of expression on introduction of miR-K12-11, we performed a Gene Set Enrichment Analysis (GSEA-P)[19] on the changes observed when comparing AcGFP-only-expressing cell populations with those expressing both AcGFP and miR-K12-11 (the signal-to-noise ratio was used as a ranking metric). We obtained a list of predefined gene sets from the Molecular Signatures Database (http://www.broad.mit.edu/gsea/msigdb/index.jsp), specifically the c3 motif gene sets, which included gene sets of predicted targets for human miRNAs[20].

To assess whether we saw a significant enrichment of hexamer or heptamer seed matches in the 3′ UTRs of genes for which we observed the most prominent changes in expression, we further performed an empirical *P*-value calculation as follows. We selected random samples of 150 genes from all probes contained in the set of our analysis defined above, and for each 10,000 of such randomized gene sets we calculated the percentage of genes that had a hexamer or heptamer seed match to miR-K12-11 in their 3′ UTR. From these 10,000 samples we obtained a null distribution, which was compared against the percentage of genes with hexamer or heptamer seed match to miR-K12-11 in the 150 most downregulated genes (ranked as above using Gene Set Enrichment Analysis). Enrichment *P* values were then determined from the null distribution as the fraction of gene sets that had an equal or higher percentage of seed matches. Histograms were created with JMP 6.0 (SAS).

Real-time quantitative PCR (qRT–PCR). Total RNA from BJAB cells was prepared with TRIzol extraction and further processed with the RNeasy Mini kit (Qiagen) including an on-column DNase digestion step (RNase-free DNase set; Qiagen). The absence of DNA contamination was proved through minus-RT controls in qRT–PCR reactions. cDNA synthesis was performed with ABI High Capacity cDNA reagents. Real-time qPCR was performed with the Quantace SybrGreen qPCR mix. qPCR was performed on an ABI7900HT machine. On each plate, four PCR replicates for 18S rRNA were run per sample to make normalization more robust. Specific target genes were measured as single data points per sample; error bars are derived from the variance of the four 18S rRNA replicates. Relative mRNA levels were calculated against 18S rRNA amplification. RPL13A (geneID 23521) was included in the analyses as a second verifying 'housekeeper' (data not shown), which however did not yield results different from those normalized against 18S rRNA. Primer sequences are given in Supplementary Table 4. For BACH1 and HIVEP2, one representative result out of two distinct primer sets per gene is shown; for BIRC4BP, one representative result for three different primer sets is shown.

Indicator assays. Each well of a 24-well dish of 293T cells was co-transfected with 2.5 ng of each indicator and control retroviral vector as well as 0.4 μg of a pNL-SIN-CMVAcGFP-based miRNA expression construct with FuGENE6 (Roche). At 48 h after transfection, Dual-Luciferase Reporter Assays (Promega) were performed as instructed. FLuc/RLuc ratios were calculated and, for co-transfection with each miRNA expression construct and empty pNL-SIN-CMV-AcGFP, values obtained for each *Fluc* construct bearing a candidate 3′ UTR were normalized on those obtained from the corresponding value of the unmodified *Fluc* vector. These normalized values were then normalized to the values derived from co-transfections with the empty pNL-SIN-CMV-AcGFP vector, which were set at 100%. Error bars were calculated from three independent experiments.

Indicator assays for miRNA activity in BJAB and BCBL1 cells (Supplementary Figs 2 and 5) were performed with the lentiviral vectors pNL-SIN-CMV-FL and pNL-SIN-CMVRL as described[17].

Antibodies and western blotting. To analyse BACH1 protein levels, cell numbers of BJAB or THP-1 cell lines were counted and equal numbers were plated in 10-cm dishes. On the following day, cells were collected by centrifugation, washed once with PBS and then lysed in denaturing lysis buffer (40 mM Tris pH 6.8, 2% sucrose, 1% SDS). Cell lysates were boiled immediately for 5 min at 95 °C and vortex-mixed. Protein concentration was determined with the BCA protein assay kit (Pierce).

To determine the abundance of Fos protein after serum starvation and induction with TPA, cells were counted and washed twice in serum-free RPMI; 5×10^5 cells ml^{-1} in serum-free RPMI were plated into wells of six-well dishes. After the cells had been starved for 26 h in serum-free medium, the medium was aspirated (BJAB cells adhere to the culture dish during serum starvation) and cells were incubated with serum-free RPMI with or without 20 ng ml^{-1} TPA for 2 h. Cells were rinsed with PBS, lysed by the addition of 200 μl of denaturing lysis buffer to each well and further processed as described above. Equal amounts of protein (about 120 μg per lane for BACH1; about 15 μg per lane for β-actin and Fos) were analysed by western blotting. Primary goat anti-BACH1 was from Santa Cruz Biotechnology (C-20, sc-14700), primary rabbit anti-Fos was a gift from T. Curran's laboratory, and primary anti-β-actin was from Santa-Cruz Biotechnology (C-4, sc-47778). Secondary anti-goat IgG horseradish peroxidase (HRP) was from Santa Cruz Biotechnology (sc-2020) and secondary anti-rabbit IgG HRP was from Amersham (NA934V). Signals were developed by using SuperSignal West Femto (Pierce) for BACH1, or Lumi-Light (Roche) in the case of β-actin and Fos.

Antagomir treatment. Antagomirs were synthesized as described[23]. Sequences were 5′- $u_sc_sggacacaggcuaagcau_su_sa_sa_s$-Chol-3′ (antagomir-miR-K12-11), 5′- $a_sc_scgccugaagucucugau_su_sa_sa_s$-Chol-3′ (antagomir-luc1309), and 5′- g_sc_suua-cacccaguuuccugu$_sa_sa_su_s$-Chol-3′ (antagomir-miR-K12-1). Lower case letters represent 2′-*O*Me-modified nucleotides, subscript 's' represents a phosphorothioate linkage, and 'Chol' represents linked cholesterol.

For the detection of Fos by western blotting, BCBL-1 cells were washed with PBS and 5×10^5 cells ml^{-1} were incubated for 25 h with serum-free medium in the presence of 1 μM antagomir. Fos expression was induced by treatment for 1 h with 20 ng ml^{-1} TPA. Cells were recovered by centrifugation, washed with PBS and lysed as described above. For indicator assays, BCBL-1 cells were washed with serum-free medium and 5×10^5 cells ml^{-1} were incubated for 4 h with serum-free medium in the presence of 1 μM antagomir. FCS was added to a final concentration of 10% and, after 1 h, cells were recovered by centrifugation and infected with mixtures of pNL-SIN-CMV-Fluc control virus and pNL-SIN-CMV-RLuc indicator viruses carrying no additional sequences or two perfect matches for miR-K12-1 or miR-K12-11 inserted 3′ to the *Rluc* ORF. In parallel, virus mixtures were also used to infect KSHV-negative BJAB cells. Indicator assays and normalization were performed as described previously[17].

LETTERS

Reconstitution of a microtubule plus-end tracking system *in vitro*

Peter Bieling[1]*, Liedewij Laan[2]*, Henry Schek[1], E. Laura Munteanu[2], Linda Sandblad[1], Marileen Dogterom[2], Damian Brunner[1] & Thomas Surrey[1]

The microtubule cytoskeleton is essential to cell morphogenesis. Growing microtubule plus ends have emerged as dynamic regulatory sites in which specialized proteins, called plus-end-binding proteins (+TIPs), bind and regulate the proper functioning of microtubules[1–4]. However, the molecular mechanism of plus-end association by +TIPs and their ability to track the growing end are not well understood. Here we report the *in vitro* reconstitution of a minimal plus-end tracking system consisting of the three fission yeast proteins Mal3, Tip1 and the kinesin Tea2. Using time-lapse total internal reflection fluorescence microscopy, we show that the EB1 homologue Mal3 has an enhanced affinity for growing microtubule end structures as opposed to the microtubule lattice. This allows it to track growing microtubule ends autonomously by an end recognition mechanism. In addition, Mal3 acts as a factor that mediates loading of the processive motor Tea2 and its cargo, the Clip170 homologue Tip1, onto the microtubule lattice. The interaction of all three proteins is required for the selective tracking of growing microtubule plus ends by both Tea2 and Tip1. Our results dissect the collective interactions of the constituents of this plus-end tracking system and show how these interactions lead to the emergence of its dynamic behaviour. We expect that such *in vitro* reconstitutions will also be essential for the mechanistic dissection of other plus-end tracking systems.

Microtubules are polar, dynamic tubulin polymers that have a variety of functions in eukaryotic cells[5]. The dynamics and the spatial organization of microtubules are regulated by several highly conserved microtubule-associated proteins. An important class of these proteins, called +TIPs, accumulates selectively at growing microtubule plus ends in living cells. A wealth of fluorescence microscopy studies in various organisms have identified numerous +TIPs that belong to conserved subfamilies: CLIP-170 (ref. 6), APC (ref. 7), EB1 (ref. 8), CLASPs (ref. 9), p150 (ref. 10) and spectraplakins[11]. In the fission yeast *Schizosaccharomyces pombe*, classical genetics combined with real-time fluorescence microscopy[12] demonstrated that multiple aspects of cellular organization depend on a defined distribution of microtubules[13,14]. This distribution is mediated by, among others, three +TIPs: the EB1 homologue Mal3 (ref. 15), the Clip170 homologue Tip1 (ref. 16) and the kinesin Tea2 (refs 17, 18). A hierarchy of molecular events required for plus-end tracking has been established from observations inside living yeast cells: the motor Tea2 and its putative cargo Tip1 move along the microtubule lattice towards its growing plus ends, where they accumulate[17,19]. Efficient recruitment to microtubules and the plus-end accumulation of Tea2 and Tip1 depend on the presence of Mal3, which itself tracks the microtubule plus ends independently of Tea2 and Tip1 (refs 15, 17, 19). It is not yet known whether additional factors or post-translational modifications are required, or whether Mal3, Tea2 and Tip1 constitute a minimal system that is sufficient to show plus-end tracking. In fact, a mechanistic understanding of plus-end tracking is still missing, in part because of the lack of an *in vitro* assay in which plus-end tracking can be reconstituted with a minimal set of pure components[20].

Here we reconstitute microtubule plus-end tracking of the three purified proteins, namely Mal3, Tea2 and Tip1, *in vitro*. We observed +TIPs and dynamic microtubules on chemically functionalized surfaces by two-colour total internal reflection fluorescence (TIRF) microscopy[21] (Fig. 1a). We first studied the three +TIPs individually and then in various combinations with fluorescently labelled and unlabelled +TIPs.

Only one of the three proteins, the EB1 homologue Mal3, was able to bind efficiently to dynamic microtubules in the absence of the others. Alexa 488-labelled Mal3 selectively accumulated at growing microtubule ends at considerable ionic strength (Fig. 1b) over a wide range of protein concentrations (Supplementary Fig. 1a). Movie sequences and the corresponding kymographs (time–space plots), revealed that Mal3 was tracking both the fast-growing plus ends and the more slowly growing minus ends (Fig. 1c). However, Mal3 did not accumulate at the ends of depolymerizing microtubules (Fig. 1c and Supplementary Movie 1) or static microtubules (Supplementary Fig. 2a). Selective tracking of free, polymerizing microtubule ends is therefore an inherent property of Mal3. Mal3–Alexa 488 also bound weakly along the entire length of microtubules (Fig. 1b, c), a behaviour that was enhanced at lower ionic strength (Supplementary Fig. 1a). This binding might reflect the previously shown preferential association of Mal3 with the lattice seam of Taxol-stabilized microtubules[22].

Two fundamentally different molecular mechanisms can be envisaged for how Mal3 accumulates at the growing microtubule end. Mal3 could co-polymerize in a complex with tubulin to the growing microtubule end, and subsequently be released. Alternatively, instead of binding to free tubulin, Mal3 could recognize a characteristic structural feature at the microtubule end. This structural feature could either be a collective property of several tubulin subunits such as the previously observed protofilament sheet[23] or a property of individual tubulin dimers that are in a GTP-bound versus a GDP-bound state[24]. To distinguish between a co-polymerization mechanism and an end-recognition mechanism, we measured the spatial distribution of Mal3 along microtubule plus ends that were growing in the presence of various tubulin concentrations but a constant Mal3 concentration.

Increased microtubule growth velocities resulting from increased tubulin concentrations led to a more comet-shaped accumulation of

[1]European Molecular Biology Laboratory, Cell Biology and Biophysics Unit, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [2]FOM Institute for Atomic and Molecular Physics (AMOLF), Kruislaan 407, 1098 SJ Amsterdam, The Netherlands.
*These authors contributed equally to this work.

Mal3–Alexa 488 at growing microtubule plus ends (Fig. 2a). Averaged fluorescence intensity profiles of Mal3–Alexa 488 comets demonstrated that after an initial peak in fluorescence the signal decreased exponentially towards the basal lattice signal (Fig. 2a). The peak fluorescence of Mal3 was largely insensitive to changes in the tubulin/Mal3 ratio (Fig. 2b). This argues against a simple co-polymerization mechanism, because such a mechanism would lead to peak signals that varied with the tubulin/Mal3 ratio. Furthermore, gel-filtration experiments showed that Mal3 does not bind to unpolymerized tubulin (Supplementary Fig. 3a). This agrees with the observation that the amount of Mal3 binding along the microtubule lattice is also independent of the tubulin concentration (Fig. 2b). Together these data support a mechanism in which Mal3 tracks microtubule ends by recognizing a structural feature.

The characteristic comet tail length obtained from exponential fits to the decays of the averaged Mal3 fluorescence intensity profiles increased linearly with the microtubule growth velocity (Fig. 2c). This suggests that microtubule ends are decorated with Mal3 for a characteristic time of about 8 s, independently of microtubule growth velocity (Fig. 2d). In contrast, the dwell time of individual Mal3–Alexa 488 molecules at growing microtubule plus ends, measured with greater temporal resolution under single-molecule imaging conditions, was only 0.282 ± 0.003 s (Fig. 2e and Supplementary Fig. 4). This indicates that individual Mal3 molecules turn over rapidly on a plus-end-specific structure that has a lifetime of about 8 s before it transforms into a normal microtubule lattice structure. A similarly fast turnover of Mal3 was also observed *in vivo*[19].

In contrast to Mal3, green fluorescent protein (GFP)-tagged Tip1 and Alexa 488-labelled Tea2 did not bind significantly to the microtubules in conditions under which selective end tracking of Mal3 was observed (Fig. 3a). Under single-molecule imaging conditions, however, rare interactions of the kinesin Tea2 with the microtubule could be observed at low ionic strengths with the use of higher frame rates. A gaussian fit to the velocity distribution yielded a mean velocity of $4.8 \pm 0.3\ \mu m\ min^{-1}$, and a single-exponential fit to the '1 − cumulative probability' distribution of the measured run lengths yielded an average run length of $0.73 \pm 0.01\ \mu m$ (Fig. 3b).

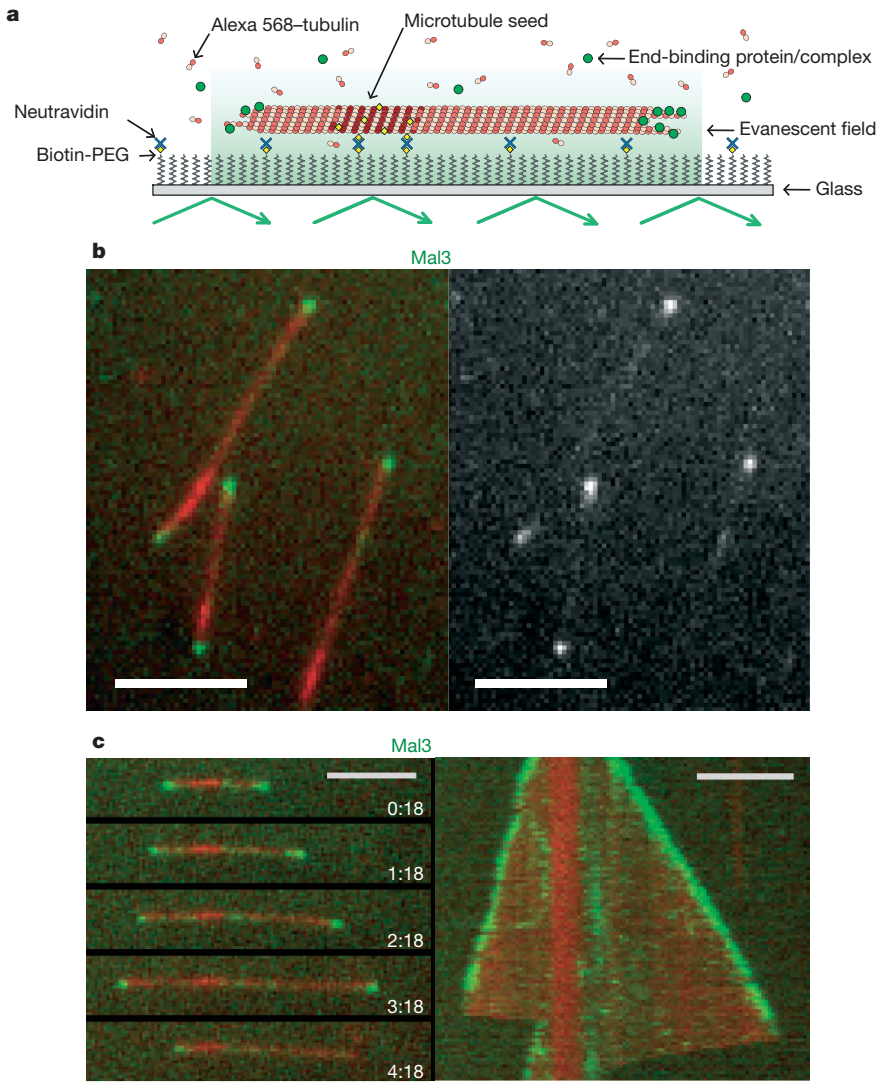


Figure 1 | Tracking of growing microtubule ends by Mal3 *in vitro*. **a**, Diagram of the experimental setup. Dynamic microtubules were grown in the presence of free Alexa 568-labelled tubulin and fluorescently labelled +TIPs from short stabilized microtubule seeds attached to a PEG-passivated glass surface by means of biotin-neutravidin links. Bright microtubule seeds, dim (non-biotinylated) microtubules extending from the seeds, and +TIPs were observed by TIRF microscopy in the evanescent field close to the glass surface. **b**, Overlaid TIRF images of Mal3–Alexa 488 (green) and dynamic Alexa 568-labelled microtubules (red) (left), and for comparison the image of Mal3–Alexa 488 alone (right). **c**, Time sequence of overlaid images of Mal3–Alexa 488 (green) and a dynamic Alexa 568-labelled microtubule (red) taken at the indicated times in minutes:seconds (left), and the corresponding kymograph of the same microtubule (right). Mal3 was used at 200 nM in all end-tracking experiments, unless otherwise stated. The kymograph displays a period of 5 min. Scale bars, 5 μm.

Because Tea2 binds *in vivo*[15] and *in vitro* (Supplementary Fig. 3b) to Tip1, and because the motor might be auto-inhibited without its putative cargo, we tested whether Tip1 could enhance the binding of Tea2–Alexa 488 to dynamic microtubules. However, this was not the case (Fig. 3a and Supplementary Movie 2).

In vivo, the presence of Mal3 is needed for plus-end tracking of Tea2 and Tip1 (refs 15, 17, 19). Using our *in vitro* approach, we examined whether the autonomous plus-end tracking protein Mal3 is sufficient to mediate microtubule plus-end tracking of the processive motor Tea2 and its cargo Tip1. In the presence of Mal3 and Tip1, Tea2–Alexa 488 now strongly accumulated at growing microtubule plus ends (Fig. 4a and Supplementary Movie 3). No accumulation of Tea2–Alexa 488 was visible at growing minus ends (Fig. 4a) or depolymerizing ends (Supplementary Fig. 1b). Furthermore, Tea2–Alexa 488 speckles appeared along the microtubule lattice and moved towards the plus end (Fig. 4a and Supplementary Movie 3). The speed of these particles was on average $9.8 \pm 2.9\,\mu m\,min^{-1}$ and therefore 4.4-fold faster than the velocity of microtubule growth ($2.2 \pm 0.3\,\mu m\,min^{-1}$; Fig. 4b). Tip1–GFP moved similarly along the microtubule lattice and also tracked growing microtubule plus ends (Fig. 4c and Supplementary Movie 4), but not depolymerizing ends (Supplementary Fig. 1b) or the ends of static microtubules (Supplementary Fig. 2b). Mal3–Alexa 488, in contrast, was not observed to move along the microtubule to the same extent as Tea2 and Tip1 (Fig. 4d and Supplementary Movie 5). These observations very closely mimic the situation *in vivo*[15,17,19].

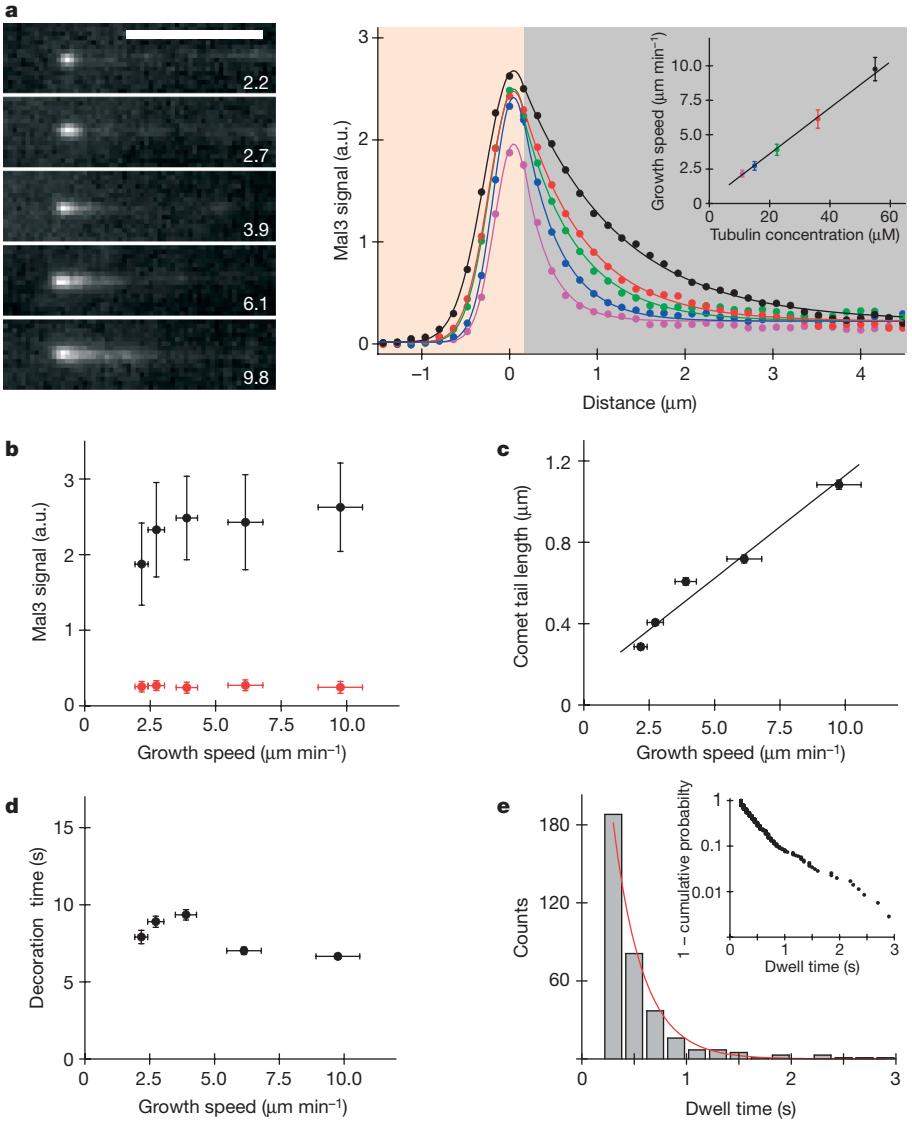


Figure 2 | Mechanism of plus-end tracking by Mal3. a, Images of individual Mal3–Alexa 488 comets at the indicated growth velocities (in $\mu m\,min^{-1}$) (left) and averaged intensity profiles of the comets (right). The Mal3–Alexa 488 concentration was 200 nM. The data (dots) were fitted (lines) using gaussian (pink area) and exponential (grey area) functions (Supplementary Methods). The inset shows the dependence of the growth velocities on tubulin concentrations. Error bars indicate s.d. **b**, The Mal3–Alexa 488 signal at the peak of the Mal3 comet (black symbols) as obtained from the averaged intensity profiles, and the signal on the microtubule lattice behind the comet (red symbols) as quantified separately from intensity line scans. Error bars indicate the s.d. of the maximum tip intensity and the s.d. of the averaged line scans for the lattice intensity. **c**, Mal3 comet tail lengths as obtained from single-exponential fits to the averaged intensity profiles. Error bars indicate standard errors as obtained from the exponential fits. **d**, The characteristic decoration time of the Mal3 signal in the Mal3 comet tail at different microtubule growth speeds as obtained by dividing the comet tail length by the microtubule growth speed. Errors were calculated by error propagation. **e**, Histogram of dwell times of single Mal3–Alexa 488 events at growing microtubule plus ends. The inset shows the '1 − cumulative probability' distribution of dwell times. The characteristic dwell time and its standard error was obtained from a fit to this distribution (red line, see Supplementary Methods). The Mal3–Alexa 488 concentration was 1 nM.

Gel filtrations demonstrated that in solution Mal3, Tea2 and Tip1 exist as a stable ternary complex (Fig. 4e). It is therefore most likely that the formation of this complex is required for efficient binding of Tea2–Tip1 to the microtubule. However, the three proteins do not behave in the same way once bound to the microtubule. Imaging the movements of the three proteins on the microtubule lattice with greater temporal resolution showed that Tip1–GFP and Alexa 647-labelled Tea2 co-migrate (Supplementary Fig. 5), indicating that Tea2 indeed transports Tip1. Consistent with this was our observation that the average run lengths for Tea2 and Tip1 were very similar, at 0.90 ± 0.01 and 1.10 ± 0.01 μm, respectively (Fig. 4f and Supplementary Fig. 6). In contrast, Mal3–Alexa 488 showed only short runs with an average run length of 0.29 ± 0.01 μm (Fig. 4f and Supplementary Fig. 6). This demonstrates that Mal3 is initially transported by Tea2, but dissociates shortly after a productive binding event.

We confirmed that Mal3-mediated recruitment of Tea2–Tip1 to the microtubule lattice requires the interaction of Mal3 with the amino-terminal extension of the kinesin Tea2 (ref. 25). Replacing full-length Tea2 with a construct lacking the N-terminal extension (ΔNTea2) abolished efficient binding of Tip1–GFP to the microtubule (Supplementary Fig. 7a). In addition, Mal3-mediated recruitment of the Tea2–Tip1 complex requires the presence of both Tea2 and Tip1. Tea2–Alexa 488 was hardly present on microtubules in the absence of Tip1 (Supplementary Fig. 7b) and Tip1–GFP was not significantly bound to microtubules in the absence of Tea2 (Supplementary Fig. 7c), whereas binding of Mal3–Alexa 488 to microtubules was unaffected in both cases (Supplementary Fig. 7d and data not shown). The results with the double combinations of proteins (Fig. 3a, right, and Supplementary Fig. 7b–d) exactly mimic the *in vivo* single-deletion mutants of *mal3*, *tea2* and *tip1* (refs 15, 17, 19).

Replacing ATP with ADP eliminated the efficient binding of Tea2–Alexa 488 along the microtubule lattice and the tracking of microtubule plus ends, despite the presence of all three proteins (Supplementary Fig. 7e and Supplementary Movie 6). Only a very weak fluorescence signal could be observed at growing microtubule ends, but without a preference for the plus or minus end (Supplementary Fig. 7e). This demonstrates that *in vitro* the processive motor activity of Tea2 is essential for efficient microtubule-end tracking of Tea2–Tip1 and also for their plus-end preference.

In living cells, single deletions of Mal3, Tea2 or Tip1 suggested that these three +TIPs mainly decrease the frequency of microtubule catastrophes without strongly affecting the other parameters of microtubule dynamic instability[15,16,18]. We tested the direct effects of Mal3 alone and of Mal3 with Tea2 and Tip1 on microtubule dynamics under conditions of selective end tracking. We imaged microtubules in the presence of unlabelled +TIPs by differential interference contrast microscopy. Similarly to the situation *in vivo*, neither Mal3 alone nor the combination of all three proteins had a strong effect on the growth and shrinkage velocities of microtubule plus ends (Supplementary Table 1). However, Mal3 alone increased the frequencies of catastrophes and rescues. The addition of Tea2–Tip1 counteracted these effects of Mal3 (Supplementary Table 1). These results show that especially the effect of Mal3 on the catastrophe frequency is different from what would be expected from the corresponding deletion *in vivo*. This is not surprising, because several other proteins not studied here are known to affect the catastrophe frequency[26,27]. By including these other modulators of microtubule dynamics in the future, our *in vitro* system promises also to lead to the identification of the more complex minimal system that reproduces physiological microtubule dynamics.

Thus, we have identified Mal3 as an autonomous tracking protein of growing microtubule ends *in vitro*. Mal3 most probably recognizes a structural feature at microtubule ends rather than co-polymerizing as a tubulin–Mal3 complex. As *in vivo*, the behaviour of Mal3 *in vitro* does not depend significantly on the presence of Tea2 or Tip1. Furthermore, we identified Mal3–Tea2–Tip1 as a minimal system producing plus-end tracking behaviour of Tea2 and Tip1 *in vitro*. This suggests that *in vivo* Tea2, Tip1 and Mal3 may also work as a

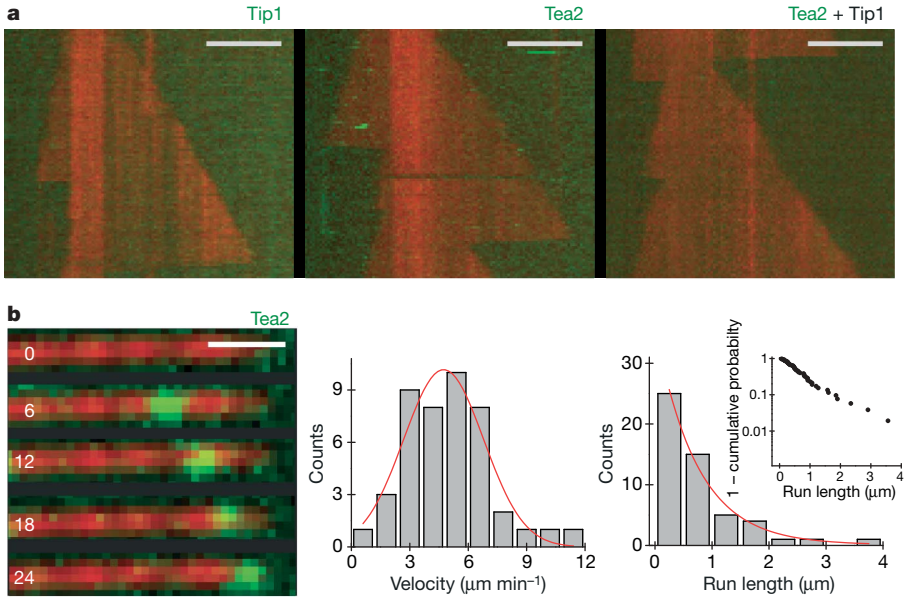


Figure 3 | Tea2 and Tip1 individually and in combination do not track microtubule ends. **a**, Kymographs of Tip1–GFP (left), Tea2–Alexa 488 (middle), and Tea2–Alexa 488 together with Tip1 (right) (labelled +TIPs in green) on dynamic Alexa 568-labelled microtubules (red). The sensitivity for GFP and Alexa 488 detection was strongly increased in comparison with that in Fig. 1b. Concentrations were 50 nM for Tip1 and 8 nM for Tea2 in all end-tracking experiments unless otherwise stated. The kymographs display a period of 5 min. Scale bars, 5 μm. **b**, Time sequence of TIRF images of a processive run of a single Tea2–Alexa 488 (see also Supplementary Fig. 8) moving on a stable Alexa 568-labelled microtubule, taken at the indicated times in seconds (left). Scale bar, 1 μm. The Tea2–GFP concentration was 0.5 nM. Histograms of velocities (centre) and run lengths (right) of single Tea2–Alexa 488 runs are shown; the inset shows the '1 – cumulative probability' distribution of run lengths. The red lines show a gaussian fit to the velocity distribution (centre) and a single-exponential fit to the '1 – cumulative probability' distribution of run lengths (right) (see Supplementary Methods).

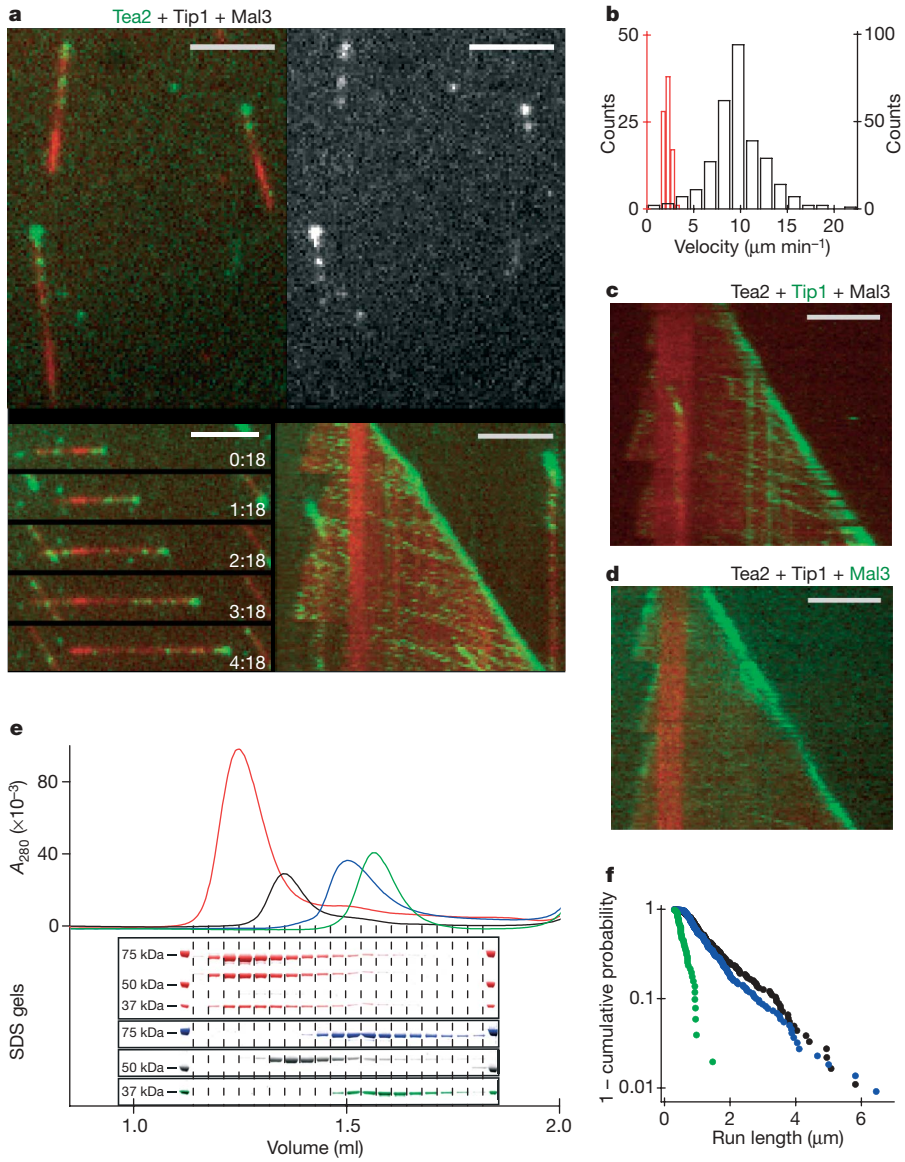


Figure 4 | Efficient microtubule plus-end tracking of Tea2–Tip1 in the presence of Mal3. a, Overlaid TIRF images showing Tea2–Alexa 488 (green) and Alexa 568-labelled microtubules (red) in the presence of the two other +TIPs (top left), and for comparison an image with the signal of only Tea2–Alexa 488 (top right). Bottom left, time sequence of images (at the times shown, in minutes:seconds); bottom right, the corresponding kymograph. Protein concentrations for Mal3 are as in Fig. 1 and for Tea2 and Tip1 as in Fig. 3a. Kymographs display periods of 5 min. Scale bars, 5 μm. **b**, Histograms of the velocities of microtubule plus-end growth (red, left axis) and Tea2–Alexa 488 speckle movement along the microtubule lattice (black, right axis). The increased velocity of Tea2 speckles in comparison with single Tea2 molecules (Fig. 3b) is mostly a consequence of an increased temperature. **c**, Kymograph showing Tip1–GFP in the presence of Tea2 and Mal3. **d**, Kymograph showing Mal3–Alexa 488 (green) in the presence of Tea2 and Tip1. The signal intensity can be directly compared with Fig. 1b. **e**, Gel filtrations of Mal3, Tea2 and Tip1: elution profiles and SDS gels of the corresponding eluted fractions of individual runs of Mal3 alone (green), Tea2 alone (blue), Tip1 alone (black) and an equimolar mixture of all three +TIPs (red). **f**, Run-length distribution of Mal3–Alexa 488 (green), Tip1–GFP (black) and Tea2–Alexa 647 (blue) moving along the microtubule lattice, always in the presence of the other two +TIPs. Concentrations were 100 nM Mal3, 50 nM Tip1 and 8 nM Tea2.

microtubule plus-end tracking system, independently of other +TIPs. However, *in vivo* part of the Mal3 pool might simultaneously function in 'parallel' end tracking systems. The role of Mal3 as a loading factor of Tea2–Tip1 involves the initial formation of a ternary complex that promotes productive encounters of Tea2–Tip1 with the microtubule lattice. Tip1 is subsequently transported by the processive motor Tea2, whereas Mal3 rapidly dissociates and is transported for only short distances.

Our *in vitro* system provides a powerful new tool to test the proposed mechanisms for microtubule end targeting of different +TIPs[20,28,29] and to analyse the interplay between plus-end tracking and the dynamic properties of microtubules that are ultimately responsible for the morphogenetic function of the microtubule cytoskeleton.

METHODS SUMMARY

Protein biochemistry. Proteins were expressed, purified and labelled as described in Supplementary Methods.

Surface chemistry. Glass coverslips were cleaned, silanized and functionalized with poly(ethylene glycol) (PEG) as described[30], and treated with *N*-hydroxysuccinimido-biotin. The biotin-PEG-functionalized slides were washed, spin-dried and stored at 4 °C. To generate passivated glass, poly(L-lysine)-PEG was dried on a glass surface and then washed extensively.

End-tracking assay. Brightly labelled, short GMP-CPP microtubules (containing 20% Alexa 568-labelled tubulin and 7.7% biotinylated tubulin) were attached by means of neutravidin to a biotin-PEG-functionalized coverslip of a flow chamber (Supplementary Methods). With the use of a custom TIRF microscopy system, dynamic microtubules and +TIPs either tagged with GFP or labelled with Alexa fluorophores were observed in the presence of 11 μM dimly labelled tubulin (containing 6.7% Alexa 568-labelled tubulin) in assay

buffer (80 mM K-PIPES pH 6.8, 85 mM KCl, 4 mM $MgCl_2$, 1 mM GTP, 1 mM EGTA, 10 mM 2-mercaptoethanol and 2 mM MgATP or MgAMP-PNP or 5 mM MgADP) containing 0.1% methylcellulose (4,000 cP; Sigma) and an oxygen scavenger system. Unless stated otherwise, we kept the final concentrations of the labelled and unlabelled +TIP proteins constant at 200 nM Mal3, 50 nM Tip1 and 8 nM Tea2. These protein concentrations were chosen after systematic variation of concentrations to allow the easy visualization of both end tracking and transport along microtubules. The temperature was 30 °C.

Data analysis. The growth trajectories of microtubules and walking tracks of Tea2–Tip1 speckles were analysed with kymographs. Single-molecule motility was analysed with kymographs and by automated particle tracking implemented in a custom software environment. To analyse the shape of Mal3 comets, line profiles of the fluorescence intensity of Mal3–Alexa 488 at growing microtubule plus ends were aligned and averaged. An exponential fit to the tail of the profile was then used to quantify the decay of the signal.

Detailed methods are described in Supplementary Methods.

Received 18 September; accepted 17 October 2007.
Published online 2 December 2007.

1. Schuyler, S. C. & Pellman, D. Microtubule 'plus-end-tracking proteins': The end is just the beginning. *Cell* **105,** 421–424 (2001).
2. Mimori-Kiyosue, Y. & Tsukita, S. 'Search-and-capture' of microtubules through plus-end-binding proteins (+TIPs). *J. Biochem.* **134,** 321–326 (2003).
3. Wittmann, T. & Desai, A. Microtubule cytoskeleton: a new twist at the end. *Curr. Biol.* **15,** R126–R129 (2005).
4. Akhmanova, A. & Hoogenraad, C. C. Microtubule plus-end-tracking proteins: mechanisms and functions. *Curr. Opin. Cell Biol.* **17,** 47–54 (2005).
5. Desai, A. & Mitchison, T. J. Microtubule polymerization dynamics. *Annu. Rev. Cell Dev. Biol.* **13,** 83–117 (1997).
6. Perez, F., Diamantopoulos, G. S., Stalder, R. & Kreis, T. E. CLIP-170 highlights growing microtubule ends *in vivo*. *Cell* **96,** 517–527 (1999).
7. Mimori-Kiyosue, Y., Shiina, N. & Tsukita, S. Adenomatous polyposis coli (APC) protein moves along microtubules and concentrates at their growing ends in epithelial cells. *J. Cell Biol.* **148,** 505–518 (2000).
8. Mimori-Kiyosue, Y., Shiina, N. & Tsukita, S. The dynamic behavior of the APC-binding protein EB1 on the distal ends of microtubules. *Curr. Biol.* **10,** 865–868 (2000).
9. Akhmanova, A. *et al.* Clasps are CLIP-115 and -170 associating proteins involved in the regional regulation of microtubule dynamics in motile fibroblasts. *Cell* **104,** 923–935 (2001).
10. Vaughan, P. S., Miura, P., Henderson, M., Byrne, B. & Vaughan, K. T. A role for regulated binding of $p150^{Glued}$ to microtubule plus ends in organelle transport. *J. Cell Biol.* **158,** 305–319 (2002).
11. Kodama, A., Karakesisoglou, I., Wong, E., Vaezi, A. & Fuchs, E. ACF7: an essential integrator of microtubule dynamics. *Cell* **115,** 343–354 (2003).
12. Ding, D. Q., Chikashige, Y., Haraguchi, T. & Hiraoka, Y. Oscillatory nuclear movement in fission yeast meiotic prophase is driven by astral microtubules, as revealed by continuous observation of chromosomes and microtubules in living cells. *J. Cell Sci.* **111,** 701–712 (1998).
13. Hayles, J. & Nurse, P. A journey into space. *Nature Rev. Mol. Cell Biol.* **2,** 647–656 (2001).
14. Brunner, D. & Nurse, P. New concepts in fission yeast morphogenesis. *Phil. Trans. R. Soc. Lond. B* **355,** 873–877 (2000).
15. Busch, K. E. & Brunner, D. The microtubule plus end-tracking proteins mal3p and tip1p cooperate for cell-end targeting of interphase microtubules. *Curr. Biol.* **14,** 548–559 (2004).
16. Brunner, D. & Nurse, P. CLIP170-like tip1p spatially organizes microtubular dynamics in fission yeast. *Cell* **102,** 695–704 (2000).
17. Browning, H., Hackney, D. D. & Nurse, P. Targeted movement of cell end factors in fission yeast. *Nature Cell Biol.* **5,** 812–818 (2003).
18. Browning, H. *et al.* Tea2p is a kinesin-like protein required to generate polarized growth in fission yeast. *J. Cell Biol.* **151,** 15–28 (2000).
19. Busch, K. E., Hayles, J., Nurse, P. & Brunner, D. Tea2p kinesin is involved in spatial microtubule organization by transporting tip1p on microtubules. *Dev. Cell* **6,** 831–843 (2004).
20. Carvalho, P., Tirnauer, J. S. & Pellman, D. Surfing on microtubule ends. *Trends Cell Biol.* **13,** 229–237 (2003).
21. Axelrod, D. Total internal reflection fluorescence microscopy in cell biology. *Traffic* **2,** 764–774 (2001).
22. Sandblad, L. *et al.* The *Schizosaccharomyces pombe* EB1 homolog Mal3p binds and stabilizes the microtubule lattice seam. *Cell* **127,** 1415–1424 (2006).
23. Chretien, D., Fuller, S. D. & Karsenti, E. Structure of growing microtubule ends: two-dimensional sheets close into tubes at variable rates. *J. Cell Biol.* **129,** 1311–1328 (1995).
24. Drechsel, D. N. & Kirschner, M. W. The minimum GTP cap required to stabilize microtubules. *Curr. Biol.* **4,** 1053–1061 (1994).
25. Browning, H. & Hackney, D. D. The EB1 homolog Mal3 stimulates the ATPase of the kinesin Tea2 by recruiting it to the microtubule. *J. Biol. Chem.* **280,** 12299–12304 (2005).
26. West, R. R., Malmstrom, T., Troxell, C. L. & McIntosh, J. R. Two related kinesins, klp5+ and klp6+, foster microtubule disassembly and are required for meiosis in fission yeast. *Mol. Biol. Cell* **12,** 3919–3932 (2001).
27. Ohkura, H., Garcia, M. A. & Toda, T. Dis1/TOG universal microtubule adaptors—one MAP for all? *J. Cell Sci.* **114,** 3805–3812 (2001).
28. Tirnauer, J. S., Grego, S., Salmon, E. D. & Mitchison, T. J. EB1–microtubule interactions in *Xenopus* egg extracts: role of EB1 in microtubule stabilization and mechanisms of targeting to microtubules. *Mol. Biol. Cell* **13,** 3614–3626 (2002).
29. Folker, E. S., Baker, B. M. & Goodson, H. V. Interactions between CLIP-170, tubulin, and microtubules: implications for the mechanism of Clip-170 plus-end tracking behavior. *Mol. Biol. Cell* **16,** 5373–5384 (2005).
30. Lata, S. & Piehler, J. Stable and functional immobilization of histidine-tagged proteins via multivalent chelator headgroups on a molecular poly(ethylene glycol) brush. *Anal. Chem.* **77,** 1096–1105 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Utz for technical assistance, protein purifications and cloning; J. Piehler for help with surface chemistry; I. Telley for help with data analysis; M. Braun and A. Seitz for helping to initiate this project; H. Besir for protein purifications; R. Santarella and S. Kandels-Lewis for cloning; G. Stier for the gift of pETM-Z; Y. Kalaidzidis and Transinsight GMBH for the gift of the PLUK MT beta version used to track moving particles; and G. Brouhard for additional help with the software. T.S. acknowledges support from the German Research Foundation (DFG), T.S. and M.D. from the European Commission (STREP Active Biomics), H.S. from EMBO, D.B. and M.D. from the Human Frontier Science Program, and M.D. from the 'Stichting voor Fundamenteel Onderzoek der Materie (FOM-NWO)'.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to T.S. (surrey@embl.de), D.B. (brunner@embl.de) or M.D. (dogterom@amolf.nl).

LETTERS

RAG2 PHD finger couples histone H3 lysine 4 trimethylation with V(D)J recombination

Adam G. W. Matthews[1]*, Alex J. Kuo[2]*, Santiago Ramón-Maiques[3], Sunmi Han[1], Karen S. Champagne[4], Dmitri Ivanov[5], Mercedes Gallardo[1], Dylan Carney[2], Peggie Cheung[2], David N. Ciccone[1], Kay L. Walter[2], Paul J. Utz[6], Yang Shi[7], Tatiana G. Kutateladze[4], Wei Yang[3], Or Gozani[2]* & Marjorie A. Oettinger[1]*

Nuclear processes such as transcription, DNA replication and recombination are dynamically regulated by chromatin structure. Eukaryotic transcription is known to be regulated by chromatin-associated proteins containing conserved protein domains that specifically recognize distinct covalent post-translational modifications on histones. However, it has been unclear whether similar mechanisms are involved in mammalian DNA recombination. Here we show that RAG2—an essential component of the RAG1/2 V(D)J recombinase, which mediates antigen-receptor gene assembly[1]—contains a plant homeodomain (PHD) finger that specifically recognizes histone H3 trimethylated at lysine 4 (H3K4me3). The high-resolution crystal structure of the mouse RAG2 PHD finger bound to H3K4me3 reveals the molecular basis of H3K4me3-recognition by RAG2. Mutations that abrogate RAG2's recognition of H3K4me3 severely impair V(D)J recombination *in vivo*. Reducing the level of H3K4me3 similarly leads to a decrease in V(D)J recombination *in vivo*. Notably, a conserved tryptophan residue (W453) that constitutes a key structural component of the K4me3-binding surface and is essential for RAG2's recognition of H3K4me3 is mutated in patients with immunodeficiency syndromes. Together, our results identify a new function for histone methylation in mammalian DNA recombination. Furthermore, our results provide the first evidence indicating that disrupting the read-out of histone modifications can cause an inherited human disease.

Many studies suggest that V(D)J recombination is regulated by modulating the chromatin structure of antigen-receptor loci during lymphoid development[2]. Several studies have analysed the pattern of histone modifications present at the immunoglobulin heavy chain locus during B-cell development[3–5]. However, mechanisms linking histone modifications to the function of the RAG recombinase have remained elusive.

Because RAG2 contains a non-canonical plant homeodomain (PHD) finger[6,7]—a module that can mediate interactions with chromatin[8–10]—we asked whether a polypeptide encompassing the RAG2 PHD finger ($RAG2_{PHD}$; amino acids 414–527) can recognize modified histone proteins. In an *in vitro* screen of peptide microarrays containing ~70 distinct modified histone peptides, we found that $RAG2_{PHD}$ specifically binds to histone H3 trimethylated at lysine 4 (H3K4me3) (Fig. 1a, Supplementary Fig. 1, and data not shown). The specificity of this interaction was confirmed by peptide pull-down assays (Fig. 1b, Supplementary Fig. 2 and Supplementary Fig. 3). RAG2 has a carboxy-terminal extension of 40 amino acids that is essential for phosphoinositide (PtdIns)-binding[7] (amino acids 488–527), but this region is dispensable for H3K4me3-binding because the minimal PHD finger alone (amino acids 414–487) is sufficient for H3K4me3-recognition (Fig. 1c). In addition, the acidic hinge region of RAG2 (amino acids 388–412), previously implicated in histone-binding[11], is dispensable for recognition of H3K4me3 (Fig. 1d). Moreover, mutations in the acidic hinge region, which had previously been shown to interfere with histone binding[11], had no effect on the H3K4me3 interaction (Fig. 1d). Using calf thymus histone protein pull-down assays, we confirmed that the interaction between $RAG2_{PHD}$ and H3K4me3 occurs in the context of full-length histone proteins (Fig. 1e). Furthermore, $RAG2_{PHD}$ bound to native mononucleosomes purified from HeLa cells, but not to mononucleosomes reconstituted from bacterially expressed recombinant histones, indicating that the $RAG2_{PHD}$–nucleosome interaction is dependent on post-translational histone modifications (Fig. 1f). Although full-length RAG2 binds H3K4me3 peptides, core RAG2 (amino acids 1–387)—the minimal portion of RAG2 required for V(D)J cleavage *in vitro*—which lacks the PHD finger, does not (Fig. 1g). Consistent with the pull-down results, tryptophan fluorescence measurements using $RAG2_{PHD}$ determined dissociation constants (K_d) of ~4 μM for H3K4me3, ~60 μM for H3K4me2, ~120 μM for H3K4me1 and ~500 μM for H3K4me0 (Supplementary Fig. 5). Thus, the RAG2 PHD finger is a chromatin-binding module that recognizes H3K4me3.

To understand the molecular basis of the interaction between the RAG2 PHD finger and H3K4me3, we determined the crystal structure of the $RAG2_{PHD}$–H3K4me3 complex at 1.15 Å resolution (Fig. 2, Supplementary Fig. 6 and Supplementary Table 1). In three previously published PHD–H3K4me3 structures[12–14], the H3 peptide extends straight through the peptide-binding groove. However, in the $RAG2_{PHD}$–H3K4me3 structure, the H3 peptide is kinked by ~90° at Q5. The K4me3 side chain is recognized by cation–π interactions within an aromatic channel delimited by Y415 on the left, M443 on the back and W453 on the right (Fig. 2a). Despite not revealing any primary sequence conservation with the K4me3-binding residues of ING2, BPTF and YNG1 (Supplementary Fig. 7), RAG2 forms a similar K4me3-binding pocket (Fig. 2c). Interestingly, $RAG2_{PHD}$ lacks the canonical 'aromatic cage' often employed to bind trimethylated lysine residues[15] (Fig. 2b, c). Instead of being closed on both sides, the back and the top (as observed with other PHD fingers[12–14]), the $RAG2_{PHD}$ K4me3-binding surface is open on the top, and resembles an 'aromatic

[1]Department of Molecular Biology, Massachusetts General Hospital and Department of Genetics, Harvard Medical School, Boston, Massachusetts 02114, USA. [2]Department of Biological Sciences, Stanford University, Stanford, California 94305, USA. [3]Laboratory of Molecular Biology, NIDDK, NIH, Bethesda, Maryland 20892, USA. [4]University of Colorado Health Sciences Center, Aurora, Colorado 80045, USA. [5]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115, USA. [6]Department of Medicine, Stanford University School of Medicine, Stanford, California 94305, USA. [7]Department of Pathology, Harvard Medical School, Boston, Massachusetts 02115, USA.
*These authors contributed equally to this work.

channel' rather than an 'aromatic cage' (Fig. 2b, c). This 'channel' conformation may provide a mechanism to modulate histone binding[16]. Aside from the K4me3 and Q5 residues, the remaining side chains of H3 form no specific interactions with $RAG2_{PHD}$. Finally, unlike other H3K4me3-binding PHD fingers[9,12–14,17], $RAG2_{PHD}$ does not have acidic residues (Asp or Glu) positioned to electrostatically interact with H3R2 (ref. 16) (Fig. 2c).

The three residues in RAG2 that form the aromatic channel critical for trimethyl-lysine recognition are completely conserved through evolution (Supplementary Fig. 8). As expected, mutating any one of these residues (Y415A, M443A, W453A/R) abrogated H3K4me3-binding by $RAG2_{PHD}$ (Fig. 3a) and by full-length RAG2 (Fig. 3b). Because the W453R mutation has been implicated in the pathogenesis of Omenn's syndrome[18]—a rare severe combined immunodeficiency[19]—and the molecular mechanism linking this mutation to Omenn's syndrome remains unknown[7], we further characterized W453R's role in histone binding. Consistent with the critical role of this residue in forming the recognition surface for H3K4me3, introducing this mutation into the RAG2 PHD finger ($RAG2_{PHD\text{-}W453R}$) abolished the ability of $RAG2_{PHD}$ to bind either full-length histone H3 (Fig. 3c) or intact nucleosomes (Fig. 3d). Thus, the interaction of RAG2 with histone proteins *in vitro* is dependent on H3K4me3-binding. We note that $RAG2_{PHD\text{-}W453R}$ is properly folded, as indicated by a comparison of the ^{1}H-^{15}N heteronuclear single quantum correlation spectra of $RAG2_{PHD}$ and $RAG2_{PHD\text{-}W453R}$ (Supplementary Fig. 9). In addition, the Y415A and M443A substitutions had no effect on PtdIns-binding by $RAG2_{PHD}$ (Supplementary Fig. 10), and W453R-associated recombination defects were previously demonstrated to be independent of the PtdIns-binding activity of RAG2 (ref. 7). Thus, the identification of multiple different point mutations that selectively abrogate RAG2's recognition of H3K4me3 allowed us to study the functional significance of this interaction.

Before analysing the effect of H3 methylation on V(D)J recombination *in vivo*, we first wanted to confirm that $RAG2_{Y415A}$, $RAG2_{M443A}$ and $RAG2_{W453R}$ disrupt H3K4me3-binding without affecting protein folding or the inherent catalytic properties of RAG2. We tested the ability of recombinant wild-type and mutant RAG2 proteins to catalyse V(D)J cleavage *in vitro* on a naked DNA substrate. In these assays, the recombinant proteins, which all expressed and purified equally well (Fig. 3e), were incubated with core RAG1 (amino acids 384−1008), and a DNA substrate. All three H3K4me3-binding mutants catalysed V(D)J cleavage at wild-type levels (Fig. 3f). Therefore, the H3K4me3-binding mutants—$RAG2_{Y415A}$, $RAG2_{M443A}$ and $RAG2_{W453R}$—are properly folded and catalytically active and, thus, can be used for *in vivo* functional analyses.

Next, to address whether the recognition of methylated H3 by $RAG2_{PHD}$ has a role in regulating RAG2 function *in vivo*, we performed extrachromosomal V(D)J recombination assays with the H3K4me3-binding mutants. Fibroblast cell lines were transfected with an exogenous recombination substrate that becomes partially chromatinized when introduced into cells and that has nucleosomes positioned over the recombination signal sequences[20], along with full-length RAG1, and either full-length wild-type RAG2, $RAG2_{Y415A}$, $RAG2_{M443A}$, or $RAG2_{W453R}$. Strikingly, despite being expressed at comparable levels to wild-type RAG2 (Fig. 4a, western blot), all three H3K4me3-binding mutants exhibited a profound decrease (>90%) in recombination activity (Fig. 4a, graph). Thus, the H3K4me3-recognition activity of RAG2 is required for V(D)J recombination *in vivo*.

To test directly the importance of H3K4 methylation and the interaction between RAG2 and H3K4me3 for V(D)J recombination *in vivo*, we used two different methods to reduce endogenous H3K4 methylation levels, and then repeated the extrachromosomal V(D)J recombination assays. First, H3K4 methylation levels were reduced (Fig. 4b, western blot) by knocking down expression of the common histone H3 lysine 4 methyltransferase component WDR5 (refs 15 and 21) with short hairpin (sh)RNA. Consistent with the decreased

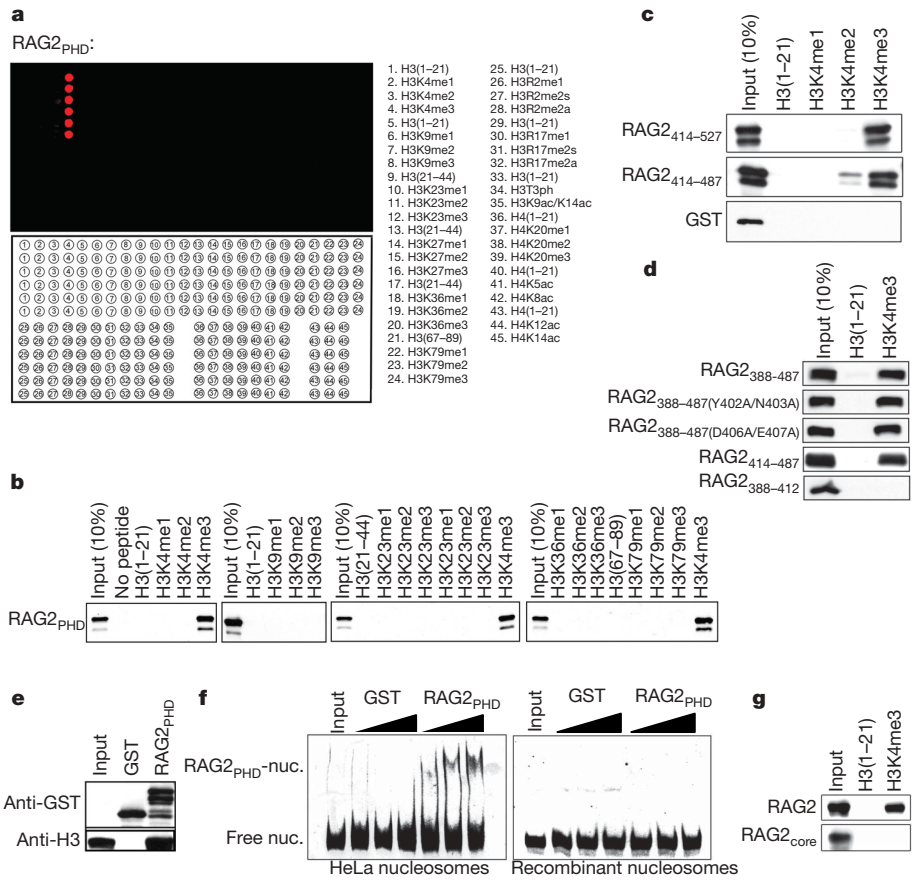


Figure 1 | The RAG2 PHD finger is a novel H3K4me3-binding module. **a**, $RAG2_{PHD}$ preferentially binds H3K4me3 peptides. Peptide microarrays containing the indicated histone peptides were probed with glutathione *S*-transferase (GST)–$RAG2_{414–527}$ ($RAG2_{PHD}$). Red spots indicate peptide binding by $RAG2_{PHD}$. The integrity of peptides was previously confirmed[9,17] (Supplementary Fig. 4). H3, histone H3; H4, histone H4; me, methylation; ac, acetylation; ph, phosphorylation; s, symmetric; a, asymmetric. **b**, Western blot analysis of histone peptide pull downs with $RAG2_{PHD}$ and indicated biotinylated peptides. **c**, $RAG2_{414–487}$ is sufficient for recognition of H3K4me3 in a histone peptide pull-down assay, as in **b**. GST alone is shown as a negative control. **d**, The RAG2 hinge region (amino acids 388–412) is dispensable for H3K4me3-recognition. Peptide pull-down assays as in **b**, with the indicated proteins. **e**, A western blot of $RAG2_{PHD}$ and GST control pull downs from calf thymus histones. **f**, A gel-shift assay comparing $RAG2_{PHD}$ binding to purified native HeLa nucleosomes (left panel) or nucleosomes reconstituted from recombinant, bacterially expressed histone proteins (right panel). Ethidium-bromide stain of nucleosomal DNA on a non-denaturing polyacrylamide gel. **g**, Full-length RAG2, but not $RAG2_{core}$ recognizes H3K4me3. Peptide pull-down assays as in **b**, with the indicated proteins.

recombination observed with the H3K4me3-binding mutants, we observed a marked reduction in the recombination activity of wild-type RAG2 (~60%) in cells carrying the *WDR5* shRNA (Fig. 4b, table and left graph). Significantly, V(D)J recombination by $RAG2_{W453R}$—which does not recognize H3K4me3—was unaffected by the presence of *WDR5* shRNA (Fig. 4b, table and right graph), indicating that the reduction observed for wild-type RAG2 is specifically due to reduced H3K4me3 binding. In a second independent method, we reduced H3K4me3 levels by transiently expressing the H3K4me3 demethylase SMCX (also known as JARID1C)[22]. As with *WDR5* shRNA, we observed that decreases in H3K4me3 levels (Fig. 4c, western blot) resulted in a significant reduction in recombination (~45%) in cells expressing SMCX (Fig. 4c, table and left graph). V(D)J recombination by $RAG2_{W453R}$ was unaffected by SMCX expression (Fig. 4c, table and right graph). Thus, reducing either the levels of H3K4 methylation or the ability of RAG2 to bind H3K4me3, impairs V(D)J recombination.

Because H3K4me3 is observed at actively rearranging gene segments[23] (Supplementary Fig. 12), we assayed the ability of the H3K4me3-binding mutants to perform chromosomal V(D)J recombination at the endogenous murine IgH locus. $RAG2^{-/-}$ Pro-B cells were transduced with lentiviruses encoding either RAG2, $RAG2_{Y415A}$, $RAG2_{M443A}$ or $RAG2_{W453R}$ and the extent of D_H-to-J_H recombination in the transduced cell populations was measured using a standard semi-quantitative PCR strategy. Despite being expressed at comparable levels to wild-type RAG2 (Fig. 4d, left panel), all three H3K4me3-binding mutants exhibited a dramatic reduction in D_H-to-J_H recombination (Fig. 4d, right panel). Thus, we conclude that recognition of H3K4me3 by $RAG2_{PHD}$ is crucial for V(D)J recombination at the endogenous immunoglobulin locus.

Our findings provide the first direct molecular link between histone methylation and mammalian DNA recombination. Before this study, H3K4me3 had only been demonstrated to function in the regulation of gene expression[15]. Our results, demonstrating a novel function for this mark in V(D)J recombination, highlight that chromatin structure can be coupled to diverse nuclear processes by protein modules that recognize modified histones. We have also provided the first evidence that disrupting the read-out of histone modifications can cause an inherited human disease. Omenn's syndrome has been observed in patients carrying a W453R mutation in RAG2 (ref. 18). Because $RAG2_{W453R}$ exhibits wild-type enzymatic activity *in vitro*[7] (Fig. 3f), it has been unclear how this mutation causes immunodeficiency. Here we have shown that the W453R mutation disrupts the read-out of H3K4me3, thereby providing a molecular explanation of how $RAG2_{W453R}$ causes Omenn's syndrome. Although the W453R,

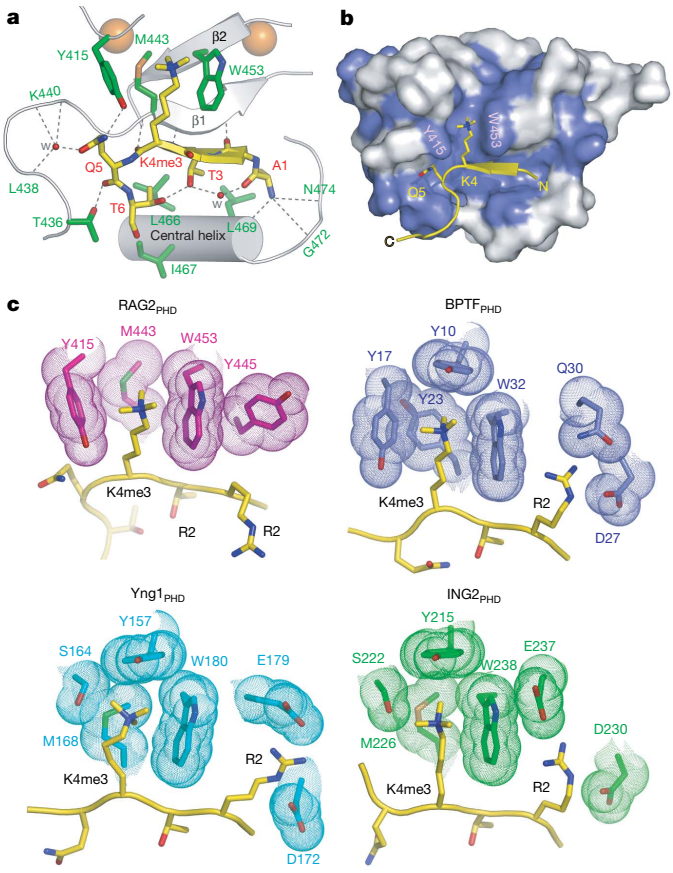


Figure 2 | The molecular basis of H3K4me3-recognition by $RAG2_{PHD}$. **a**, **b**, A 1.15 Å crystal structure of $RAG2_{414–487}$ complexed with H3K4me3 peptide. **a**, Ribbon diagram of the complex. For clarity, only the central portion of $RAG2_{414–487}$ is shown (silver). Residues with side chains that interact with the H3 peptide are shown as green sticks with blue (nitrogen) and red (oxygen) highlights. Residues with main chain atoms that interact with the peptide are labelled in green. The peptide is shown in yellow. With the exception of R2, the side chains of A1 to T6 all interact with RAG2 and are shown in stick models. Zn^{2+} ions are shown as orange spheres and water molecules mediating protein–peptide interactions are shown as red spheres. Grey dashed lines represent hydrogen bonds. $RAG2_{PHD}$ consists of two unorthodox, interdigitated zinc fingers, linked by a pair of anti-parallel β-strands and a central α-helix. The backbone of residues two–four of the histone H3 peptide are hydrogen-bonded, with one of the β-strands of $RAG2_{PHD}$ forming a 3-stranded antiparallel β-sheet. **b**, The H3 peptide binding surface is conserved among RAG2 proteins. Residues of RAG2 that are conserved through evolution (see Supplementary Fig. 5) are coloured blue on the molecular surface. **c**, Structural comparison of the PHD fingers from murine RAG2, human BPTF, yeast Yng1 and murine ING2. Side chains in the PHD fingers that interact with H3K4me3 and H3R2 are highlighted with molecular surface.

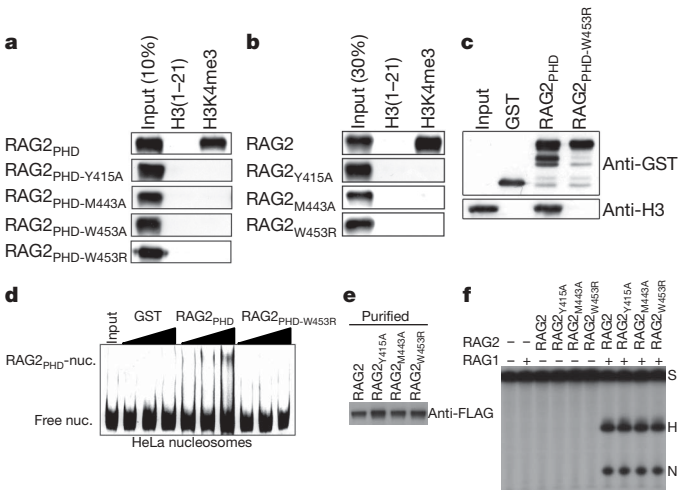
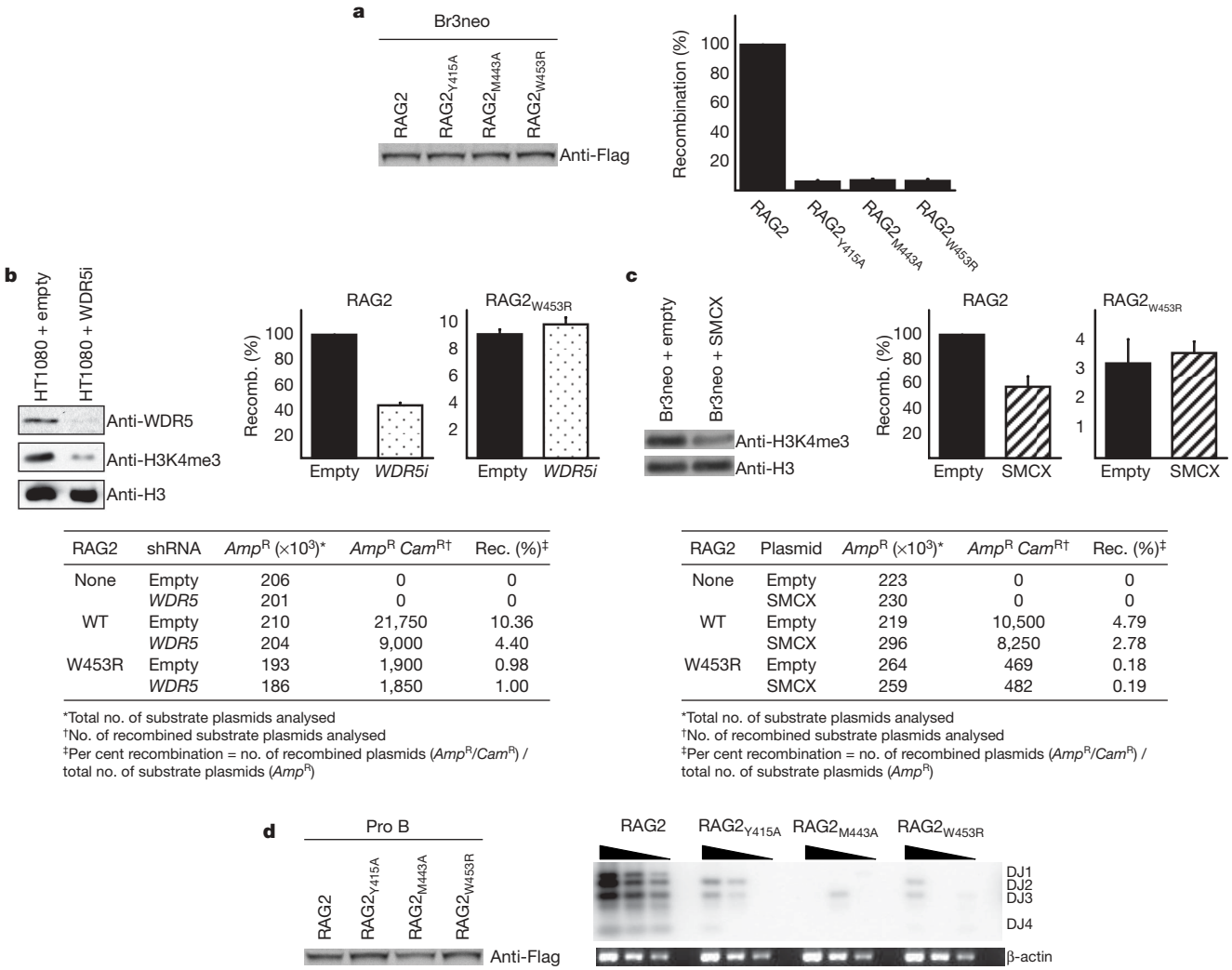


Figure 3 | Recognition of H3K4me3 by $RAG2_{PHD}$ is dispensable for RAG2 *in vitro* enzymatic activity, but essential for RAG2 binding to native histones. **a**, Identification of RAG2 PHD finger mutations that specifically disrupt H3K4me3-recognition. Western blot analysis of histone peptide pull downs, with the indicated GST fusion proteins and biotinylated peptides. **b**, $RAG2_{PHD}$ mutations specifically disrupt binding of full-length RAG2 to H3K4me3. Western blot analysis of histone peptide pull downs with wild-type and mutant full-length RAG2 proteins and the indicated biotinylated histone peptides. **c**, The interaction of RAG2 with histone H3 is dependent on H3K4me3 binding. Western analysis of GST–$RAG2_{PHD}$, GST–$RAG2_{PHD-W453R}$, and GST control pull downs from calf thymus histones. **d**, The interaction of RAG2 with native nucleosomes requires H3K4me3-binding activity. Nucleosome-binding assays as in Fig. 1f with wild-type ($RAG2_{PHD}$) and mutant ($RAG2_{PHD–W453R}$) GST-fusion proteins. **e**, Wild-type and mutant RAG2 proteins express and purify equally well. Anti-Flag western analysis of the indicated Flag-tagged full-length RAG2 proteins purified from 293T cells. **f**, Mutant RAG2 proteins catalyse wild-type V(D)J cleavage *in vitro*. The indicated recombinant proteins were tested for *in vitro* V(D)J cleavage activity. The positions of the substrate (S) and cleavage products (hairpin (H) and nick (N)) are indicated.



RAG2	shRNA	*Amp*^{R} (×10^3)*	*Amp*^{R} *Cam*^{R}†	Rec. (%)‡
None	Empty	206	0	0
	WDR5	201	0	0
WT	Empty	210	21,750	10.36
	WDR5	204	9,000	4.40
W453R	Empty	193	1,900	0.98
	WDR5	186	1,850	1.00

*Total no. of substrate plasmids analysed
†No. of recombined substrate plasmids analysed
‡Per cent recombination = no. of recombined plasmids (*Amp*^{R}/*Cam*^{R}) / total no. of substrate plasmids (*Amp*^{R})

RAG2	Plasmid	*Amp*^{R} (×10^3)*	*Amp*^{R} *Cam*^{R}†	Rec. (%)‡
None	Empty	223	0	0
	SMCX	230	0	0
WT	Empty	219	10,500	4.79
	SMCX	296	8,250	2.78
W453R	Empty	264	469	0.18
	SMCX	259	482	0.19

*Total no. of substrate plasmids analysed
†No. of recombined substrate plasmids analysed
‡Per cent recombination = no. of recombined plasmids (*Amp*^{R}/*Cam*^{R}) / total no. of substrate plasmids (*Amp*^{R})

Figure 4 | Recognition of H3K4me3 is crucial for RAG1/2 recombinase activity *in vivo*. **a**, The H3K4me3-recognition activity of RAG2 is required for extrachromosomal V(D)J recombination *in vivo*. Left panel, western analysis of Flag-tagged full-length RAG2 proteins expressed in Br3neo human fibroblast cells confirms that wild-type (WT) and mutant RAG2 proteins are expressed at comparable levels *in vivo*. Right panel, The indicated constructs were used for transient V(D)J recombination assays in Br3neo cells. Recombination activity was normalized to wild-type activity (RAG2), which was defined as 100%. Wild-type RAG2 consistently gave recombination frequencies of ~5–6%. Results represent the mean ±s.d. of six independent experiments. **b**, Reducing H3K4 methylation levels specifically impairs V(D)J recombination by wild-type RAG2. Western analysis demonstrates that the *WDR5* shRNA vector (WDR5i) reduces H3K4me3 levels. Transient V(D)J recombination assays performed with wild-type (left graph) or mutant (right graph) RAG2 in HT1080 human fibroblast cells in the presence (stippled) or absence (filled) of a *WDR5* shRNA vector. Results represent the mean ±s.d. of six independent experiments. Table, representative recombination data. **c**, Reducing H3K4me3 levels by demethylation specifically reduces V(D)J recombination by wild-type RAG2. Western analysis confirms that SMCX expression reduces H3K4me3 levels. Transient V(D)J recombination assays were performed with wild-type (left graph) or mutant (right graph) RAG2 in Br3neo cells in the presence (hatched) or absence (filled) of SMCX. Results represent the mean ±s.d. of six independent experiments. Table, representative recombination data. **d**, The H3K4me3-recognition activity of RAG2 is required for chromosomal V(D)J recombination at the endogenous murine IgH locus. Left panel, western analysis of Flag-tagged full-length RAG2 proteins expressed in lentivirally transduced *RAG2*$^{-/-}$ Pro-B cells confirms that wild-type and mutant RAG2 proteins are expressed at comparable levels *in vivo*. Right panel, endogenous V(D)J recombination assays in Pro-B cells transduced as in the western blot. Southern blot analysis of PCR-amplified genomic DNA on an agarose gel. Serial dilutions represent PCR-amplification of 400, 200 and 100 ng of genomic DNA. Results shown are representative of four independent experiments.

M443A, and Y415A point mutations—which selectively disrupt H3K4me3-recognition—severely disrupt V(D)J recombination, complete deletion of the RAG2 C terminus (including the PHD finger) only partially compromises V(D)J recombination activity[24–27] (Supplementary Fig. 11). We propose that the RAG2–H3K4me3 interaction serves two functions: (1) increases the stable association of RAG1/2 complexes at target sites and (2) might relieve an inhibitory activity present in the C-terminal portion of RAG2 (ref. 7) (Supplementary Fig. 13). Taken together with other studies that have demonstrated a link between the writing of histone modifications and human disease[15,28,29], our results highlight the fundamental role chromatin has in human health. We postulate that, in the coming years, other naturally occurring mutations that cause inherited human diseases will be linked to disruption of lysine methylation signalling pathways.

Note added in proof: While this work was under review, another study also reported that the RAG2 PHD finger binds to methylated H3K4 (ref. 30).

METHODS SUMMARY

Biotinylated peptides were synthesized at Stanford or Yale Protein and Nucleic Acid facilities. Peptide microarray experiments were performed essentially as described[17]. Biotinylated peptide pull-down assays, calf thymus histone binding assays, and mononucleosome gel-shift assays were performed as

described[9]. Tryptophan fluorescence assays were performed essentially as described[13]. Nuclear magnetic resonance (NMR) structure determination was performed as described[7]. The structure determination is described in Methods. *In vitro* V(D)J cleavage assays were performed as described[31]. The extrachromosomal V(D)J recombination assays and the endogenous V(D)J recombination assay are described in Methods. Information about antibodies is available in Methods.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 24 July; accepted 2 November 2007.
Published online 21 November 2007.

1. Gellert, M. V(D)J recombination: RAG proteins, repair factors, and regulation. *Annu. Rev. Biochem.* **71,** 101–132 (2002).
2. Oettinger, M. A. How to keep V(D)J recombination under control. *Immunol. Rev.* **200,** 165–181 (2004).
3. Chowdhury, D. & Sen, R. Stepwise activation of the immunoglobulin μ heavy chain gene locus. *EMBO J.* **20,** 6394–6403 (2001).
4. Johnson, K., Angelin-Duclos, C., Park, S. & Calame, K. L. Changes in histone acetylation are associated with differences in accessibility of V_H gene segments to V-DJ recombination during B-cell ontogeny and development. *Mol. Cell. Biol.* **23,** 2438–2450 (2003).
5. Morshead, K. B., Ciccone, D. N., Taverna, S. D., Allis, C. D. & Oettinger, M. A. Antigen receptor loci poised for V(D)J rearrangement are broadly associated with BRG1 and flanked by peaks of histone H3 dimethylated at lysine 4. *Proc. Natl Acad. Sci. USA* **100,** 11577–11582 (2003).
6. Callebaut, I. & Mornon, J. P. The V(D)J recombination activating protein RAG2 consists of a six-bladed propeller and a PHD fingerlike domain, as revealed by sequence analysis. *Cell. Mol. Life Sci.* **54,** 880–891 (1998).
7. Elkin, S. K. *et al.* A PHD finger motif in the C terminus of RAG2 modulates recombination activity. *J. Biol. Chem.* **280,** 28701–28710 (2005).
8. Ragvin, A. *et al.* Nucleosome binding by the bromodomain and PHD finger of the transcriptional cofactor p300. *J. Mol. Biol.* **337,** 773–788 (2004).
9. Shi, X. *et al.* ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression. *Nature* **442,** 96–99 (2006).
10. Wysocka, J. *et al.* A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature* **442,** 86–90 (2006).
11. West, K. L. *et al.* A direct interaction between the RAG2 C terminus and the core histones is required for efficient V(D)J recombination. *Immunity* **23,** 203–212 (2005).
12. Li, H. *et al.* Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. *Nature* **442,** 91–95 (2006).
13. Pena, P. V. *et al.* Molecular mechanism of histone H3K4me3 recognition by plant homeodomain of ING2. *Nature* **442,** 100–103 (2006).
14. Taverna, S. D. *et al.* Yng1 PHD finger binding to H3 trimethylated at K4 promotes NuA3 HAT activity at K14 of H3 and transcription at a subset of targeted ORFs. *Mol. Cell* **24,** 785–796 (2006).
15. Ruthenburg, A. J., Allis, C. D. & Wysocka, J. Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark. *Mol. Cell* **25,** 15–30 (2007).
16. Ramon-Maiques, S. *et al.* The PHD finger of RAG2 recognizes histone H3 methylated at both lysine-4 and arginine-2. *Proc. Natl Acad. Sci. USA*. doi:10.1073/pnas.0709170104 (in the press).
17. Shi, X. *et al.* Proteome-wide analysis in *Saccharomyces cerevisiae* identifies several PHD fingers as novel direct and selective binding modules of histone H3 methylated at either lysine 4 or lysine 36. *J. Biol. Chem.* **282,** 2450–2455 (2007).
18. Gomez, C. A. *et al.* Mutations in conserved regions of the predicted RAG2 kelch repeats block initiation of V(D)J recombination and result in primary immunodeficiencies. *Mol. Cell. Biol.* **20,** 5653–5664 (2000).
19. Villa, A. *et al.* Partial V(D)J recombination activity leads to Omenn syndrome. *Cell* **93,** 885–896 (1998).
20. Baumann, M., Mamais, A., McBlane, F., Xiao, H. & Boyes, J. Regulation of V(D)J recombination by nucleosome positioning at recombination signal sequences. *EMBO J.* **22,** 5197–5207 (2003).
21. Sims, R. J. III & Reinberg, D. Histone H3 Lys 4 methylation: caught in a bind? *Genes Dev.* **20,** 2779–2786 (2006).
22. Iwase, S. *et al.* The X-linked mental retardation gene *SMCX/JARID1C* defines a family of histone H3 lysine 4 demethylases. *Cell* **128,** 1077–1088 (2007).
23. Perkins, E. J., Kee, B. L. & Ramsden, D. A. Histone 3 lysine 4 methylation during the pre-B to immature B-cell transition. *Nucleic Acids Res.* **32,** 1942–1947 (2004).
24. Akamatsu, Y. *et al.* Deletion of the RAG2 C terminus leads to impaired lymphoid development in mice. *Proc. Natl Acad. Sci. USA* **100,** 1209–1214 (2003).
25. Corneo, B. *et al.* Rag mutations reveal robust alternative end joining. *Nature* **449,** 483–486 (2007).
26. Kirch, S. A., Rathbun, G. A. & Oettinger, M. A. Dual role of RAG2 in V(D)J recombination: catalysis and regulation of ordered Ig gene assembly. *EMBO J.* **17,** 4881–4886 (1998).
27. Liang, H. E. *et al.* The ''dispensable'' portion of RAG2 is necessary for efficient V-to-DJ rearrangement during B and T cell development. *Immunity* **17,** 639–651 (2002).
28. Shi, Y. & Whetstine, J. R. Dynamic regulation of histone lysine methylation by demethylases. *Mol. Cell* **25,** 1–14 (2007).
29. Tenney, K. & Shilatifard, A. A. COMPASS in the voyage of defining the role of trithorax/MLL-containing complexes: linking leukemogensis to covalent modifications of chromatin. *J. Cell. Biochem.* **95,** 429–436 (2005).
30. Liu, Y., Subrahmanyam, R., Chakraborty, T., Sen, R. & Desiderio, S. A plant homeodomain in Rag-2 that binds hypermethylated lysine 4 of histone H3 is necessary for efficient antigen-receptor-gene rearrangement. *Immunity* **4,** 561–571 (2007).
31. Matthews, A. G., Elkin, S. K. & Oettinger, M. A. Ordered DNA release and target capture in RAG transposition. *EMBO J.* **23,** 1198–1206 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank K.-J. Armache and J.-J. Song for the generous gift of recombinant mononucleosomes; R. Kingston and members of the Kingston laboratory for helpful discussions; and N. Lau, J.-J. Song, and M. Gellert for critical reading of this manuscript. This work was supported by NIH grants (M.A.O., O.G., D.I. and T.G.K.), as well as a Korea Research Foundation grant (S.H.). O.G. is a recipient of a Burroughs Wellcome Career Award in Biomedical Sciences and a Kimmel Scholar Award. S.R.-M. has been the recipient of a fellowship from the Human Frontier Science Program. A.J.K. is funded by Stanford University through a Genentech Foundation Predoctoral Fellowship. A.G.W.M. is a Howard Hughes Medical Institute Predoctoral Fellow.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.A.O. (oettinger@frodo.mgh.harvard.edu) or O.G. (ogozani@stanford.edu).

METHODS

Materials and plasmids. Antibodies used in the study were: anti-Histone H3 (Abcam); anti-Flag M5 (Sigma); anti-Flag M2 (Sigma); anti-glutathione *S*-transferase (GST; Santa Cruz); donkey anti-rabbit IgG HRP-linked F(ab')$_2$ fragment (GE Healthcare); sheep anti-mouse IgG HRP-linked whole antibody (GE Healthcare); and Alexa Fluor 647 chicken anti-rabbit IgG (Invitrogen). GST–RAG2$_{PHD}$ (amino acids 414–487), GST–RAG2$_{PHD}$ (amino acids 414–527) and Flag-tagged RAG2 were described previously[7]. Point mutations were generated by site-directed mutagenesis PCR using Pfu turbo DNA polymerase (Stratagene). The GST–PHD finger of hING2 (amino acids 200–281) was described previously[9].

Peptide microarray. Peptide microarray experiments were performed as described previously[17]. Briefly, biotinylated histone peptides were printed in six replicates onto a streptavidin-coated slide (ArrayIt) using a VersArray Compact Microarrayer (Bio-Rad). After a short blocking incubation with biotin (Sigma), the slides were incubated with GST-fused RAG2$_{PHD}$ in peptide binding buffer (50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 0.1% Nonidet P-40, 20% fetal bovine serum) overnight at 4 °C with gentle agitation. After washing with the same buffer, slides were probed first with anti-GST antibody and then fluorescein-conjugated secondary antibody and visualized with a GenePix 4000 scanner (Molecular Devices).

Biotinylated-peptide binding assays. Biotinylated-peptide pull-down assays were performed as described previously[9]. Briefly, 1 μg of biotinylated peptides were incubated with 1 μg of GST–PHD-fingers in peptide binding buffer (50 mM Tris-HCl, pH 7.5, 300 mM NaCl, 0.1% Nonidet P-40) overnight at 4 °C. After 1 h incubation with streptavidin beads (Amersham), complexes were washed 3 times with the binding buffer, and the bound proteins were subjected to either western or Coomassie analysis.

Calf thymus histone binding assays. Calf thymus histone binding assays were performed as described previously[9]. Briefly, 10 μg of GST–RAG2$_{PHD}$ was incubated with 25 μg of calf thymus histones (Worthington) in binding buffer (50 mM Tris-Hcl, pH 7.5, 1 M NaCl, 1% Nonidet P-40) overnight at 4 °C. After being incubated with glutathione beads for 1 h, complexes were washed 3 times with binding buffer, and bound proteins were subjected to western analysis.

Mononucleosome shift assay. Mononucleosome gel-shift assays were performed as described previously[9]. Briefly, 1.5 μg of mononucleosomes isolated from HeLa cells or assembled from recombinant histones[32] were incubated with GST–RAG2$_{PHD}$ in binding buffer (20 mM HEPES, pH 7.9, 80 mM KCl, 0.1 mM, $ZnCl_2$, 0.1% EDTA and 10% glycerol) at 30 °C for 30 min. The reaction was then subjected to 5% TBE native gel electrophoresis and visualized with ethidium-bromide staining.

Tryptophan fluorescence spectroscopy. The fluorescence spectra were recorded at 25 °C on a Fluoromax3 spectrofluorometer. The samples of 10 μM RAG2$_{PHD}$ containing progressively increasing concentrations (up to 2 mM) of histone H3 peptides (amino acids 1–12) were excited at 295 nm. Emission spectra were recorded between 305 and 405 nm with a 0.5-nm step size and a 1-s integration time and averaged over 3 scans. The K_ds were determined by a non-linear least-squares analysis using the equation: $\Delta I = (\Delta I_{max} * L)/(K_D + L)$, where L is concentration of the histone peptide, ΔI is observed change of signal intensity, and ΔI_{max} is the difference in signal intensity of the free and bound states of the RAG2 PHD finger. The K_d value was averaged over two experiments for the H3K4me3 binding and over three separate experiments for the binding of H3K4me0, H3K4me1 and H3K4me2 peptides.

Data collection and structure determination. The RAG2 PHD finger (amino acids 414–487) was prepared as described[16]. Crystals of its complex with H3K4me3 were obtained at 3 mg ml^{-1} protein concentration and 1:1.5 molar ratio of protein to peptide with a precipitant solution of 26% PEG 3350 and 0.18 M potassium thiocyanate. X-ray diffraction data were collected from a single crystal of RAG2$_{PHD}$– H3K4me3 complex on a Mar225 charge-coupled device detector at ID-22 1 Å beamline in Advanced Photon Source (APS) at −160 °C. The crystal belongs to the C2 space group with 2 RAG2$_{PHD}$–H3K4me3 complexes per asymmetric unit (Supplementary Table 1). Ramachandran statistics: 93.3% allowed region, 6.7% additionally allowed region. The data set was processed and scaled at 1.15 Å using HKL2000 (ref. 33). Crystallographic phases were obtained by molecular replacement with PHASER[34], using as a model the 2.4 Å resolution structure of RAG2$_{PHD}$–H3K4me3 determined previously[16]. The model was traced using COOT[35] and refined using CNS[36] and SHELX[37]. Individual anisotropic displacement parameters were refined for all atoms, and hydrogens were added in the late stages of refinement.

***In vitro* V(D)J cleavage assays.** Flag-tagged RAG2 and derivatives were used for *in vitro* V(D)J cleavage assays as described previously[38]. Briefly, V(D)J cleavage assays were initiated by the addition of R1 (~80 ng) and R2 (~10 ng) proteins to a 10 μl reaction mixture containing 0.25 pmol of ^{32}P-labelled, uncleaved 12-RSS substrate (VDJ100/101) in 60 mM K-glutamate, 1 mM $MnCl_2$, 25 mM Hepes, pH 7.5, and 2mM DTT. After a 2 h incubation at 30 °C, the reactions were stopped by addition of 95% formamide loading dye. The samples were denatured by heating at 95 °C for 5 min, and reaction products were visualized by autoradiography of samples separated by denaturing electrophoresis.

Extrachromosomal V(D)J recombination assays. Extrachromosomal V(D)J recombination assays were performed as described previously[39] in either Br3neo[39] or HT1080[9] human fibroblastoid cells, and the plasmid pGG49 was used as the reporter[40]. Full-length RAG1 was transiently expressed using the pcDNA6-myc-hisA vector (Invitrogen). RAG2 and derivatives were transiently expressed using the p3xFLAG-CMV vector (Sigma). Expression of all proteins was confirmed by western analysis. For the *WDR5* shRNA experiments, HT1080 cells were first stably transfected with either a *WDR5* shRNA vector or a control vector that lacks the shRNA insert. These cells were subsequently transfected with *RAG1*, *RAG2* and a recombination substrate. Activity was normalized to that of wild-type RAG2 in the absence of WDR5, which was defined as 100%. For the SMCX experiments, Br3neo cells were simultaneously co-transfected with *RAG1*, *RAG2*, a recombination substrate, and either an *SMCX* mammalian expression vector, or a control vector that lacks the *SMCX* insert. Activity was normalized to that of wild-type RAG2 in the absence of SMCX, which was defined as 100%.

Endogenous V(D)J recombination assays. Endogenous V(D)J recombination assays were performed essentially as described previously[26,41], except that *RAG2*$^{-/-}$ Pro B cells were lentivirally transduced with *RAG2* and derivatives. Expression of all proteins was confirmed by western analysis.

32. Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**, 251–260 (1997).
33. Ottinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
34. McCoy, A. J. Solving structures of protein complexes by molecular replacement with Phaser. *Acta Crystallogr. D Biol. Crystallogr.* **63**, 32–41 (2007).
35. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
36. Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
37. Sheldrick, G. M. & Schneider, T. R. SHELXL: High-resolution refinement. *Methods Enzymol.* **277**, 319–343 (1997).
38. Elkin, S. K., Matthews, A. G. & Oettinger, M. A. The C-terminal portion of RAG2 protects against transposition *in vitro*. *EMBO J.* **22**, 1931–1938 (2003).
39. Dai, Y. *et al.* Nonhomologous end joining and V(D)J recombination require an additional factor. *Proc. Natl Acad. Sci. USA* **100**, 2462–2467 (2003).
40. Gauss, G. H. & Lieber, M. R. Unequal signal and coding joint formation in human V(D)J recombination. *Mol. Cell. Biol.* **13**, 3900–3906 (1993).
41. Schlissel, M. S., Corcoran, L. M. & Baltimore, D. Virus-transformed pre-B cells show ordered activation but not inactivation of immunoglobulin gene rearrangement and transcription. *J. Exp. Med.* **173**, 711–720 (1991).

LETTERS

Crystal structure of the plasma membrane proton pump

Bjørn P. Pedersen[1,2]*, Morten J. Buch-Pedersen[1,2,3]*, J. Preben Morth[1,2], Michael G. Palmgren[1,3] & Poul Nissen[1,2]

A prerequisite for life is the ability to maintain electrochemical imbalances across biomembranes. In all eukaryotes the plasma membrane potential and secondary transport systems are energized by the activity of P-type ATPase membrane proteins: H^+-ATPase (the proton pump) in plants and fungi[1–3], and Na^+,K^+-ATPase (the sodium–potassium pump) in animals[4]. The name P-type derives from the fact that these proteins exploit a phosphorylated reaction cycle intermediate of ATP hydrolysis[5]. The plasma membrane proton pumps belong to the type III P-type ATPase subfamily, whereas Na^+,K^+-ATPase and Ca^{2+}-ATPase are type II[6]. Electron microscopy has revealed the overall shape of proton pumps[7], however, an atomic structure has been lacking. Here we present the first structure of a P-type proton pump determined by X-ray crystallography. Ten transmembrane helices and three cytoplasmic domains define the functional unit of ATP-coupled proton transport across the plasma membrane, and the structure is locked in a functional state not previously observed in P-type ATPases. The transmembrane domain reveals a large cavity, which is likely to be filled with water, located near the middle of the membrane plane where it is lined by conserved hydrophilic and charged residues. Proton transport against a high membrane potential is readily explained by this structural arrangement.

The *Arabidopsis thaliana* auto-inhibited H^+-ATPase 2 (AHA2) is a well-characterized member of the plasma membrane proton pump family[8]. As shown in Fig. 1, we have determined the structure of an active form of AHA2, devoid of a flexible, carboxy-terminal regulatory domain (R domain)[3,9] and in complex with adenosine 5′-(β,γ-methylene)-triphosphate (AMPPCP, a non-hydrolysable ATP analogue). Despite anisotropy of the data, we successfully traced the structure and refined a model encompassing residues 12 to 844 and the bound nucleotide on the basis of experimental electron-density maps calculated at 3.6 Å resolution (Fig. 2a, b, Supplementary Table 1 and Supplementary Fig. 1).

The structure of AHA2 consists of four clearly defined domains: a transmembrane domain with ten helices, M1 through M10, and three cytosolic domains, named after their counterparts in the Ca^{2+}-ATPase[10] as N (nucleotide binding; residues 338–488), P (phosphorylation; residues 308–337 and 489–625) and A (actuator; residues 12–57 and 129–233). AHA2 and the rabbit SERCA1a Ca^{2+}-ATPase share low sequence homology (20% identity; Supplementary Fig. 2), but a structural comparison shows the overall fold to be remarkably similar, supporting the assumption that the overall structure of P-type ATPases is conserved among different subfamilies (Fig. 2c, d). However, the N domain of AHA2 is smaller than the N domains found in the type II subfamily. It has the same fold as the N domain of the *Archaeoglobus fulgidus* copper pump[11], although the loops connecting strand 3 to strand 4 and strand 5 to strand 6 are longer and resemble the loops found in Ca^{2+}-ATPase[10] and Na^+,K^+-ATPase[12]. AMPPCP is bound with the adenosine part at the N domain and the triphosphate group protruding towards the P domain (Figs 1 and 2a). The N domain is inserted into the P domain through a hinge (including the conserved sequence motif DPPR) and with bound nucleotide it can move towards the P domain

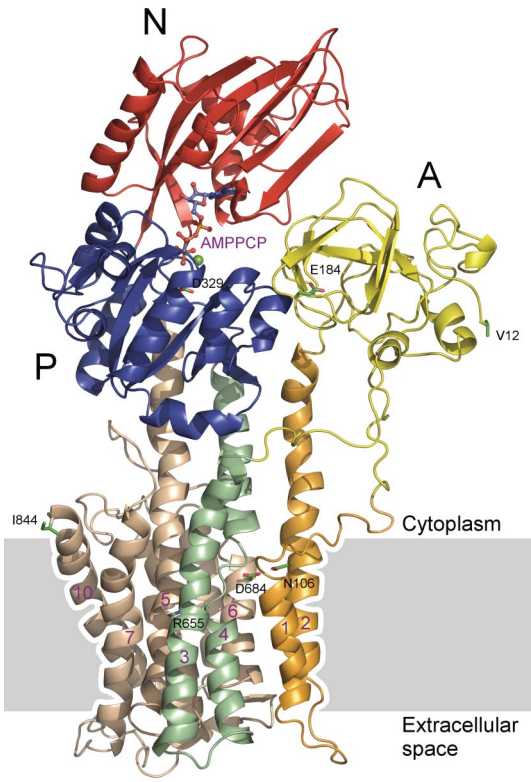


Figure 1 | Overall structure of the plasma membrane H^+-ATPase. The structure represents an active form of the proton pump, without its auto-inhibitory C terminus, in complex with Mg-AMPPCP. Ten transmembrane helices, orange, green and brown, as indicated; nucleotide-binding domain (N), red; the phosphorylation domain (P), blue; and the actuator domain (A), yellow. Mg-AMPPCP is found at the interface between the N and P domains and is shown as a ball-and-stick reprentation. Key residues mentioned in the text are shown as sticks. The grey box depicts the approximate location of the plasma membrane.

[1]Centre for Membrane Pumps in Cells and Disease—PUMPKIN. Danish National Research Foundation, [2]Department of Molecular Biology, University of Aarhus, Gustav Wieds Vej 10C, DK-8000 Aarhus, Denmark. [3]Plant Physiology and Anatomy Laboratory, Department of Plant Biology, University of Copenhagen, Thorvaldsensvej 40, DK-1871 Frederiksberg, Denmark.
*These authors contributed equally to this work.

to assemble the catalytic site, at which Asp 329 will become phosphorylated once every pumping cycle. The A domain, which stimulates dephosphorylation of Asp 329, is situated on top of M2, which protrudes as a pole out of the membrane, and it is further connected to the M1 and M3 transmembrane segments by extended loops. Glu 184 in the conserved motif TGES, which is involved in the A domain phosphatase functionality, is situated ~28 Å from Asp 329. This affirms that a large rotation of the A domain towards the P domain is required for dephosphorylation to occur, linking events at the phosphorylation site to conformational changes in the membrane. In the transmembrane domain, the M1 helix shows a prominent 90° kink (Fig. 2b) imposed by a proline residue, Pro 68, that is conserved in type III P-type ATPases[6]. A similar kink is seen in the Ca^{2+}-ATPase[13] and Na^+,K^+-ATPase[12], despite distinct motifs in the M1 primary structure for each type. M4 is unwound in the middle of the transmembrane segment, and M7 and M10 are tilted approximately 25° and 45°, respectively, relative to the plane of the membrane.

The overall arrangement of domains and transmembrane helices of AHA2 is similar, but not identical, to the occluded E1 form of Ca^{2+}-ATPase trapped in the transition state of phosphoryl transfer[14,15] (Fig. 2d). The A domain is moved away from the P domain,

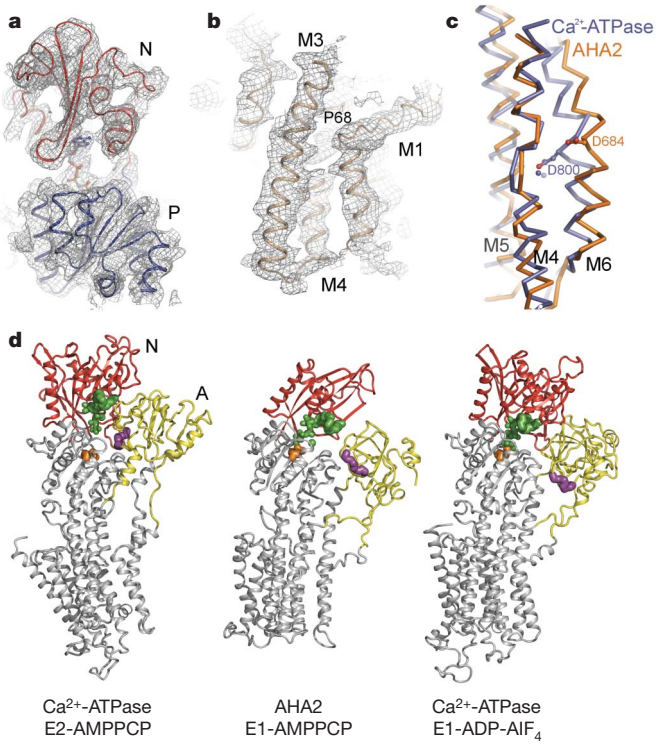


Figure 2 | Structural conservation of P-type ATPase architecture. a, View of the N (red) and P (blue) domains of the H^+-ATPase with bound Mg-AMPPCP together with the experimental electron-density map contoured at 1σ. The adenosine of the nucleotide is bound at the N domain, whereas the triphosphate group and the magnesium ion extend towards the P domain. **b**, View of the transmembrane region with the experimental electron-density map contoured at 1σ. M1 exhibits a ~90° kink perpendicular to the membrane face facilitated by Pro 68. **c**, AHA2 (orange) is aligned to the Ca^{2+}-ATPase (blue, PDB code 1T5T) on transmembrane segments M4 and M5. The bulge at Asp 684 is clearly visible. **d**, Structural comparison between AHA2 and Ca^{2+}-ATPase indicates that the H^+-ATPase structure represents a novel E1 intermediate. Middle panel, AHA2 H^+-ATPase with bound Mg-AMPPCP (this study); left panel, an E2 form of the Ca^{2+}-ATPase without bound calcium (PDB code 2C8K); right panel, a Ca^{2+}-occluded E1 form of the Ca^{2+}-ATPase[15,16] in the transition state of phosphoryl transfer (PDB code 1T5T). The structures are aligned by their P domains. N domain, red; bound nucleotide, green; A-domain, yellow; TGES motif required for dephosphorylation, magenta; phosphorylation site, orange.

allowing the N domain to approach as required for phosphorylation to occur, but closure of the active site at the interface between the N and P domains has not completed. Further comparison to the AMPPCP-bound E2 form of Ca^{2+}-ATPase[16] (Fig. 2d) indicates that our AHA2 structure indeed represents a new E1 intermediate, which is compact, yet not completely occluded.

Auto-inhibition by C-terminal regulatory R domains is characteristic of type III P-type H^+-ATPases. We have obtained crystals and collected a 5.5 Å resolution data set of full-length AHA2 in a detergent-activated form. We observe additional electron density for approximately 13 residues (modelled as a helix) of the R domain, extending from the M10 helix towards a large solvent channel in the crystal (Supplementary Fig. 3). However, we do not observe density for the bulk of the R domain (residue 858 to 948), indicating that it has no defined structure in the active form of the protein. If we plot residues that, when mutated, inhibit R domain interaction (shown by mutagenesis of plant and fungal P-type proton pumps[2,17,18]), a pattern emerges in which the R domain may attain inhibition by winding around the body of the pump to interact with the A domain and the top of the M1 and M2 segments. In this position, the R domain potentially blocks entry of protons to the transmembrane binding site and restricts A domain rotations that are essential for functional transitions in the pumping cycle. This is much like the fixation by thapsigargin of the transmembrane domain in Ca^{2+}-ATPase[14], and possibly similar to the effect of regulatory peptides like sarcolipin and phospholamban[19].

Asp 684, conserved in all plasma membrane H^+-ATPases, is the only acidic residue buried in the transmembrane domain of AHA2 (Fig. 3a). Mutational studies have shown this residue to be essential for proton transport and E1–E2 transitions, and thus the most likely candidate for the proton-binding site of P-type H^+-ATPases[9,20]. Asp 684 corresponds to the essential Ca^{2+}-coordinating residue Asp 800 in the Ca^{2+}-ATPase and it is situated in M6, next to a large cavity in the membrane (Fig. 2c, see below). Asp 684 is juxtaposed to the completely conserved Asn 106 of M2 (Fig. 3b, c), which is compatible with hydrogen bonding between the two. This feature suggests an elegant coupling mechanism of H^+-ATPase between formation of the phosphorylation site in the cytoplasmic domains and occlusion of the proton-binding site with the protonated Asp 684 and Asn 106 pair buried between the M2, M4 and M6 segments. This will also readily explain the proton specificity of H^+-ATPase; the specificity arises at the protonated Asp 684–Asn 106 pair, which serves as the 'gate keeper' along the transport pathway.

Owing to a conserved Pro 286 residue, M4 is unwound, which leads to exposure of backbone carbonyl and amide groups from residues 282–286 to a large cavity in the middle of the membrane (Fig. 3). The conserved residues Tyr 645, Tyr 648, Thr 653, Arg 655 (all in M5) and Asn 683 (in M6) expose their charged or polar side chains to this cavity. The corresponding residues in M5 of the yeast PMA1 H^+-ATPase (Tyr 691, Tyr 694, Ser 699 and His 701) have been shown to be essential for proton pumping[21]. The cavity is defined by M4, M5 and M6 (Fig. 3), and in dimensions it is substantially bigger than the Ca^{2+}-binding sites I and II of the Ca^{2+}-ATPase[10]. The enlargement is mainly due to M6, which is bulged at Asp 684 (Fig. 2c, and Supplementary Fig. 4). The cavity is large enough (~380 $Å^3$) to accommodate about 12 water molecules. Proton access from the cytoplasm to the proton-binding site seems nearly closed in our structure, but could occur through an entrance pathway located between M1, M2 and M4 (Fig. 3a). Several conserved H^+-ATPase residues are positioned in this area of the pump, for example, Asn 106, Glu 113, Glu 114 (all in M2), and they could be involved in proton transfer to Asp 684 at the edge of the cavity.

Arg 655 of M5 is situated in the middle of the membrane domain, at the cavity opposite to the Asp 684 residue (Fig. 3). Arg 655 is important, but not indispensable, for proton transport[20]. The interaction of Arg 655 with the, presumably water-filled, cavity is well defined, even though the exact side-chain structure is not at the given

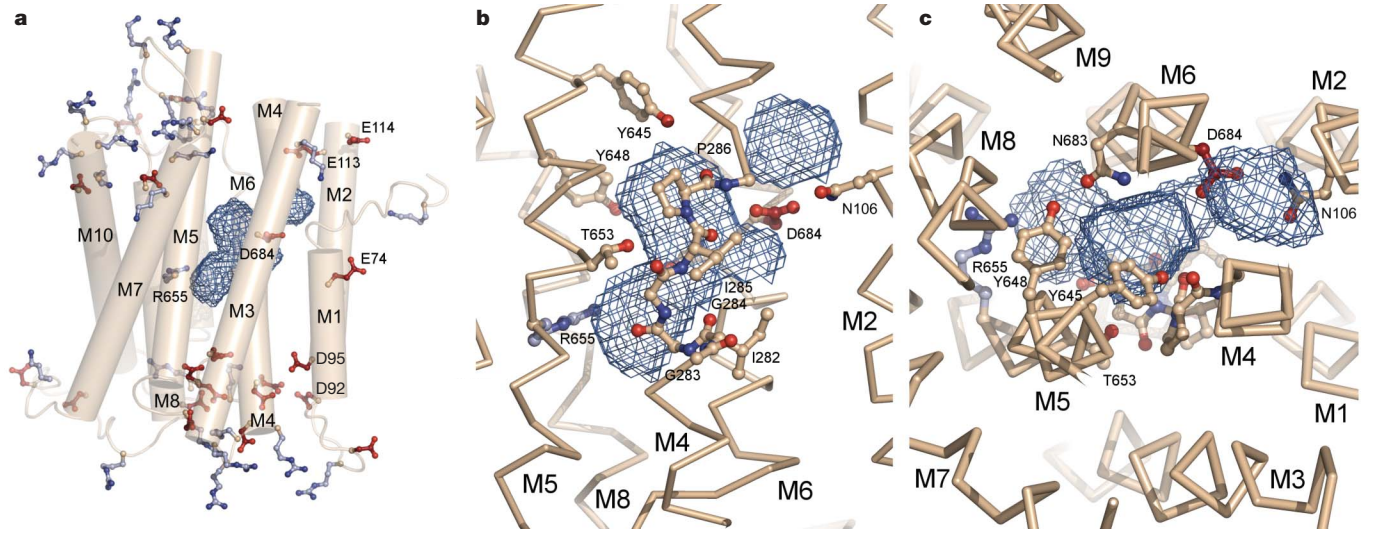


Figure 3 | The intramembranous buried cavity and proton binding site of the plasma membrane H^+-ATPase. **a**, Distribution of charged residues (Arg/Lys, blue; Asp/Glu, red) in the transmembrane region of the pump shown together with the identified intramembranous cavity (blue mesh). Among the charged residues in this region, Arg 655 and Asp 684 are the only two charged residues found in the transmembrane part (except Glu 74 of M1, which points towards the cytosolic interface). **b**, Side view and **c**, top view of the cavity and the residues lining it. Polar and charged residues from M5 and M6 together with exposed backbone carbonyls and amide groups from M4 define the boundaries of the cavity. A small extension (~80 Å^3) of the cavity, placed over Asp 684 and Asn 106, is located along the expected proton entrance pathway and is likely to close as the phosphorylation site is fully assembled.

resolution. Owing to the packing of nearby membrane residues, Arg 655 is confined to side-chain rotamer configurations pointing upwards towards Asp 684 and in direct contact with the cavity. The cavity may aid in delocalization of the buried positive charge on Arg 655 as the pump goes through phosphorylation and it may form the upper part of the proton exit pathway during subsequent proton release. The electrostatic field of Arg 655 is likely to influence Asp 684. A spatial arrangement of an arginine residue placed near an essential proton donor/acceptor is well characterized in unrelated proton pumps like bacteriorhodopsin[22,23] and F-/V-type ATPases[24,25] as a means of stimulating proton release, and we find it likely that a similar role is achieved here. Arg 655 must be expected to impose an effect on Asp 684 in the E2P state, in which the proton exit pathway opens to the extracellular environment. Also, the presence of similarly conserved positive–negative amino acid pairs of M5/M6 in H^+, K^+-ATPases (for example, Lys 800 and Asp 833 of the human ATPase ATP12A) at equivalent positions hints at a conserved mechanism of proton transport in proton-exporting P-type ATPases. Proton release from the pump might be aided by conserved acidic residues (for example, Asp 92 and Asp 95) found at the extracellular side (Fig. 3a).

Arg 655 may serve other important roles. The corresponding residue in Ca^{2+}-ATPase is Glu 771, which becomes exposed at the bottom of the Ca^{2+} exit pathway in the E2P form[26]. Likewise, Arg 655 may become exposed in the proton exit pathway of AHA2. In this position, it could serve as a positive plug that prevents proton reflux to the transmembrane binding site (Fig. 4). A positive charge along the transport pathway may then explain how proton pumps are able to generate high membrane potentials. In plants, membrane potentials may exceed −200 mV (negative on the inside; ref. 27), whereas the proton pumps from fungi generate the highest known membrane potentials (up to −300 mV; ref. 28). In fungal H^+-ATPases, Arg 655 of AHA2 is replaced by a histidine (His 701 in yeast PMA1), whereas another arginine is found at the 649 position of the M5 helix, also facing the water-filled cavity. This arrangement of an arginine and a histidine positioned in tandem at the upper part of the proton exit pathway is indeed compatible with the even-higher resistance to membrane potential that is attained in fungal proton pumps. No counter transport during the E2 to E1 transition has been described for proton pumps; this is in contrast to other subfamilies of P-type ATPases. Arg 655 may act as a built-in counter-ion during dephosphorylation and E2 to E1 transition, neutralizing the deprotonated negatively charged Asp 684. The presence of Arg 655 as a constitutive counter-ion makes the transition from E2-P to E2 extremely favourable and minimizes exposure of Asp 684 to the extracellular side. This is consistent with the fact that proton pumps cannot be directly phosphorylated by inorganic phosphate, contrary to other P-type

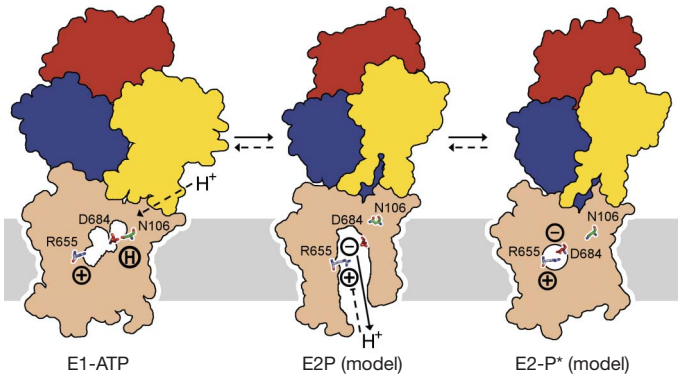


Figure 4 | Mechanism of proton transport by plasma membrane H^+-ATPase. E2-model forms of the H^+-ATPase were made by structural alignment of our E1-AMPPCP structure with the E2P structure of the Ca^{2+}-ATPase[26] and the E2-P* structure (E2 occluded state of the pump) of the Ca^{2+}-ATPase (PDB code 1XP5). Asp 684 is the central proton donor/acceptor of the pump, and together with Arg 655 it lines a centrally located water-filled cavity. In the E1 conformation, hydrogen bonding between Asp 684 and Asn 106 gives preference to the protonated form of Asp 684 (E1-ATP structure). Conformational movements in the membrane region, coupled to E1–E2 transitions, result in opening of the cavity towards the proton exit pathway (E2P model) and interrupt hydrogen bonding between Asn 106 and Asp 684; this results in proton release from Asp 684, now exposed to the extracellular environment. Placement of Arg 655 towards Asp 684 at the exit channel also stimulates proton release from Asp 684, and provides a positively charged plug in this area of the molecule that prevents extracellular protons from re-protonating Asp 684. At the same time Arg 655 functions as a built-in counter-ion that neutralizes the negative charge on Asp 684 and promotes swift formation of the occluded E2-P* transition state (E2P* model), dephosphorylation and transition to the E2 form.

ATPase subfamilies[29]. The price of being able to sustain a high membrane potential may thus be the loss of counter transport. Indeed the apparent stoichiometry of P-type H^+-ATPase transport is one proton per ATP hydrolysed[30].

The structure described here contributes to further understanding of the structural/functional relationships found in plant and fungal P-type H^+-ATPases and furthermore provides a framework for new studies of members of this subfamily. Our observation of an Asp–Asn pair and an arginine residue lining a water-filled cavity in the membrane represent key elements of proton transport by proton P-type ATPases and a novel use of the P-type ATPase architecture for active transport.

METHODS SUMMARY

Expression and purification was based on a *Saccharomyces cerevisiae* expression system. Solubilization and purification were performed with dodecyl-maltoside (DDM) as the detergent and the purified protein was dialysed against a buffer containing octaethylene glycol monododecyl ether ($C_{12}E_8$) and 5-cyclohexyl-1-pentyl-β-D-maltoside (CYMAL-5) detergents. Crystals were obtained using polyethylene glycol 400 as the precipitant. Crystals were cryoprotected by a controlled dehydration procedure by vapour diffusion, which also improved diffraction properties. Crystallographic data were collected at the beam line X06SA of the Swiss Light Source (SLS). Phases were determined using derivative crystals with $HoCl_3$, K_2PtCl_6 and Ta_6Br_{12}, respectively. Heavy-atom-derived phases were refined and extended to the maximum resolution of the native data by density modification, exploiting two-fold rotational non-crystallographic symmetry, a solvent content of 75% and several data sets showing a low level of isomorphism for inter-crystal averaging. The experimental electron density was of high quality, showing continuous backbone density but lacking detail owing to anisotropy and low resolution of the data (Fig. 2, and Supplementary Fig. 4). Final refinement using data extending to 3.6 Å resolution produced a model with a crystallographic *R*-factor of 35.0% and a free *R*-factor of 36.6%.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 20 August; accepted 26 October 2007.

1. Serrano, R., Kielland-Brandt, M. C. & Fink, G. R. Yeast plasma membrane ATPase is essential for growth and has homology with ($Na^+ + K^+$), K^+- and Ca^{2+}-ATPases. *Nature* **319,** 689–693 (1986).
2. Morsomme, P., Slayman, C. W. & Goffeau, A. Mutagenic study of the structure, function and biogenesis of the yeast plasma membrane H^+-ATPase. *Biochim. Biophys. Acta* **1469,** 133–157 (2000).
3. Palmgren, M. G. Plant plasma membrane H^+-ATPases: Powerhouses for nutrient uptake. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **52,** 817–845 (2001).
4. Skou, J. C. & Esmann, M. The Na,K-ATPase. *J. Bioenerg. Biomembr.* **24,** 249–261 (1992).
5. Pedersen, P. & Carafoli, E. Ion motive ATPases. 1. Ubiquity, properties, and significance to cell function. *Trends Biochem. Sci.* **12,** 146–150 (1987).
6. Axelsen, K. B. & Palmgren, M. G. Evolution of substrate specificities in the P-type ATPase superfamily. *J. Mol. Evol.* **46,** 84–101 (1998).
7. Auer, M., Scarborough, G. A. & Kuhlbrandt, W. Three-dimensional map of the plasma membrane H^+-ATPase in the open conformation. *Nature* **392,** 840–843 (1998).
8. Harper, J. F., Manney, L., DeWitt, N. D., Yoo, M. H. & Sussman, M. R. The *Arabidopsis thaliana* plasma membrane H^+-ATPase multigene family. Genomic sequence and expression of a third isoform. *J. Biol. Chem.* **265,** 13601–13608 (1990).
9. Buch-Pedersen, M. J., Venema, K., Serrano, R. & Palmgren, M. G. Abolishment of proton pumping and accumulation in the E1P conformational state of a plant plasma membrane H^+-ATPase by substitution of a conserved aspartyl residue in transmembrane segment 6. *J. Biol. Chem.* **275,** 39167–39173 (2000).
10. Toyoshima, C., Nakasako, M., Nomura, H. & Ogawa, H. Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 Å resolution. *Nature* **405,** 647–655 (2000).
11. Sazinsky, M. H., Mandal, A. K., Arguello, J. M. & Rosenzweig, A. C. Structure of the ATP binding domain from the *Archaeoglobus fulgidus* Cu^+-ATPase. *J. Biol. Chem.* **281,** 11161–11166 (2006).
12. Morth, J. P. *et al.* Crystal structure of the sodium–potassium pump. *Nature* doi:10.1038/nature06419 (this issue).
13. Toyoshima, C. & Nomura, H. Structural changes in the calcium pump accompanying the dissociation of calcium. *Nature* **418,** 605–611 (2002).
14. Sorensen, T. L., Moller, J. V. & Nissen, P. Phosphoryl transfer and calcium ion occlusion in the calcium pump. *Science* **304,** 1672–1675 (2004).
15. Toyoshima, C. & Mizutani, T. Crystal structure of the calcium pump with a bound ATP analogue. *Nature* **430,** 529–535 (2004).
16. Jensen, A. M., Sørensen, T. L., Olesen, C., Møller, J. V. & Nissen, P. Modulatory and catalytic modes of ATP binding by the calcium pump. *EMBO J.* **25,** 2305–2314 (2006).
17. Eraso, P. & Portillo, F. Molecular mechanism of regulation of yeast plasma membrane H^+-ATPase by glucose. Interaction between domains and identification of new regulatory sites. *J. Biol. Chem.* **269,** 10393–10399 (1994).
18. Morsomme, P., Dambly, S., Maudoux, O. & Boutry, M. Single point mutations distributed in 10 soluble and membrane regions of the *Nicotiana plumbaginifolia* plasma membrane PMA2 H^+-ATPase activate the enzyme and modify the structure of the C-terminal region. *J. Biol. Chem.* **273,** 34837–34842 (1998).
19. MacLennan, D. H., Abu-Abed, M. & Kang, C. Structure–function relationships in Ca^{2+} cycling proteins. *J. Mol. Cell. Cardiol.* **34,** 897–918 (2002).
20. Buch-Pedersen, M. J. & Palmgren, M. G. Conserved Asp684 in transmembrane segment M6 of the plant plasma membrane P-type proton pump AHA2 is a molecular determinant of proton translocation. *J. Biol. Chem.* **278,** 17845–17851 (2003).
21. Dutra, M. B., Ambesi, A. & Slayman, C. W. Structure–function relationships in membrane segment 5 of the yeast Pma1 H^+-ATPase. *J. Biol. Chem.* **273,** 17411–17417 (1998).
22. Pebay-Peyroula, E., Rummel, G., Rosenbusch, J. P. & Landau, E. M. X-ray structure of bacteriorhodopsin at 2.5 angstroms from microcrystals grown in lipidic cubic phases. *Science* **277,** 1676–1681 (1997).
23. Luecke, H., Richter, H. T. & Lanyi, J. K. Proton transfer pathways in bacteriorhodopsin at 2.3 angstrom resolution. *Science* **280,** 1934–1937 (1998).
24. Hutcheon, M. L., Duncan, T. M., Ngai, H. & Cross, R. L. Energy-driven subunit rotation at the interface between subunit a and the c oligomer in the F_O sector of *Escherichia coli* ATP synthase. *Proc. Natl Acad. Sci. USA* **98,** 8519–8524 (2001).
25. Fillingame, R. H. & Dmitriev, O. Y. Structural model of the transmembrane F_O rotary sector of H^+-transporting ATP synthase derived by solution NMR and intersubunit cross-linking *in situ*. *Biochim. Biophys. Acta* **1565,** 232–245 (2002).
26. Olesen, C. *et al.* The structural basis of calcium transport by the calcium pump. *Nature* doi:10.1038/nature06418 (this issue).
27. Hirsch, R. E., Lewis, B. D., Spalding, E. P. & Sussman, M. R. A role for the AKT1 potassium channel in plant nutrition. *Science* **280,** 918–921 (1998).
28. Blatt, M. R., Rodriguez-Navarro, A. & Slayman, C. L. Potassium–proton symport in *Neurospora*: Kinetic control by pH and membrane potential. *J. Membr. Biol.* **98,** 169–189 (1987).
29. Amory, A., Goffeau, A., McIntosh, D. B. & Boyer, P. D. Exchange of oxygen between phosphate and water catalyzed by the plasma membrane ATPase from the yeast *Schizosaccharomyces pombe*. *J. Biol. Chem.* **257,** 12509–12516 (1982).
30. Briskin, D. P. & Reynolds-Niesman, I. Determination of H^+/ATP stoichiometry for the plasma membrane H^+-ATPase from red beet (*Beta vulgaris* L.) storage tissue. *Plant Physiol.* **95,** 242–250 (1991).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank T. L.-M. Sørensen for contributions to initial project design and E. Pohl, C. Schulze-Briese and T. Tomizaki for assistance with synchrotron data collection. We are grateful to A. M. Nielsen for technical assistance and L. Yatime for help with data collection. B.P.P. is supported by a PhD fellowship from the Graduate School of Science at the University of Aarhus, M.J.B-P. by a post-doctoral fellowship from the Carlsberg Foundation, J.P.M. by a post-doctoral fellowship from the DANSYNC programme of the Danish Research Council, and P.N. by a Hallas-Møller stipend from the Novo Nordisk Foundation.

Author Contributions B.P.P. performed crystallization experiments, collected and processed the data, and determined, refined and analysed the structure. M.J.B.-P. performed expression and protein purification, and later performed crystallization experiments and analysed the structure. J.P.M. assisted in data collection and processing. M.G.P. and P.N. contributed with equal resources, supervised the project and analysed the structure.

Author Information Coordinates and structure factors have been deposited in the Protein Data Bank with the accession number 3B8C. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to P.N. (pn@mb.au.dk) and M.G.P. (palmgren@life.ku.dk).

METHODS

Sample preparation. A *Saccharomyces cerevisiae* expression construct contained nucleotides coding for a C-terminal truncated version of the AHA2 protein lacking the last 73 residues[9]. The construct includes a MRGSH6 C-terminal tag. The construct encoding wild-type protein contained nucleotides coding for amino acid 1 to 948, including the same tag. Transformed yeast were grown and harvested essentially as described[9]. Yeast were resuspended in 50 mM Mes-KOH, pH 6.5, 26% (v/v) glycerol, 50 mM KCl, 10 mM EDTA, 1 mM dithiothreitol (DTT), 1.2 mM ATP, 0.3 mM phenylmethylsulfonyl fluoride (PMSF) and 3 μg ml^{-1} pepstatin A before being broken mechanically using glass beads at 4 °C. After cell breakage, the homogenate was centrifuged for 5 min at 1,400*g*. Centrifugation of the supernatant (15 min, 12,000*g*) was followed by sedimentation of microsomal membranes by ultracentrifugation at 50,000 r.p.m. for 1 h (Beckman 70Ti rotor). An additional ultracentrifugation step at 50,000 r.p.m. for 1 h (Beckman 70Ti rotor) after resuspension of the pellet in GMEKD (50 mM Mes, pH 6.5), 20% (v/v) glycerol, 1 mM EDTA, 1 mM DTT, 50 mM KCl) supplemented with 0.2 mM PMSF and 2 μg ml^{-1} pepstatin A, allowed harvesting and homogenization of the total membrane fraction in the same buffer. Membrane proteins were solubilized at 10 mg ml^{-1} using DDM at a detergent to protein ratio of 3:1 (w/w) in 50 mM Mes-KOH, pH 6.5, 20% (v/v) glycerol, 50 mM KCl, 0.7 mM DTT and 0.7 mM EDTA. Solubilization was performed with gentle stirring for 30 min, after which unsolubilized material was removed by ultracentrifugation for 1 h at 30,000 r.p.m. (Beckman 70Ti rotor). Solubilized protein was diluted with 1 volume of 50 mM Mes, pH6.5, 20% (v/v) glycerol, 500 mM KCl, 20 mM imidazole, 0.15% (w/v) DDM including 6–8 ml Ni-NTA resin pre-equilibrated in the same buffer. PMSF and pepstatin A were added to final concentrations of 0.2 mM and 2 μg ml^{-1}, respectively, and, following batch binding for 16 h, the resin was washed with 30 volumes of wash buffer (50 mM Mes-KOH, pH 6.5, 20% (v/v) glycerol, 5 mM imidazole, 0.15% (w/v) DDM, 0.5 mM EDTA, 0.5 mM DTT, 0.2 mM PMSF, 2 μg ml^{-1} pepstatin) supplemented with 500 mM KCl, with 20 volumes of wash buffer with 250 mM KCl, and 20 volumes of wash buffer with 50 mM KCl before bound protein was eluted with 50 mM Mes-KOH, pH 6.5, 20% (v/v) glycerol, 200 mM imidazole, 0.04% (w/v) DDM, 50 mM KCl 0.5 mM EDTA, 0.5 mM DTT, 0.2 mM PMSF and 2 μg ml^{-1} pepstatin A. Eluted protein were dialysed against GMEKD and concentrated to 20–30 mg ml^{-1} on spin columns. Before crystallization experiments, protein was dialysed overnight against 50 mM KCl, 50 mM Mes pH 6.5, 10% sucrose (w/v), 1 mM DTT, 0.09 mM (critical micelle concentration) octaethyleneglycol mono-*n*-dodecylether ($C_{12}E_8$) and 2.4 mM (critical micelle concentration) 5-cyclohexyl-1-pentyl-β-D-maltoside (Cymal-5). After dialysis, 5 mM AMPPCP and 15 mM $MgCl_2$ were added and a final ultracentrifugation spin (70,000 r.p.m., 15 min) was applied before the crystallization setup.

Crystal Growth. Crystals were grown at 4 °C using the vapour diffusion method in 4 μl hanging drops with a reservoir containing 29–32% (w/v) PEG 400, 100 mM KCl, 100 mM Mes, pH 6.0, and 5% sucrose. Crystals, with a final size of around $100 \times 100 \times 200$ μm^3 obtained after typically two weeks crystal growth, were dehydrated by step-wise increase of the PEG 400 concentration in the reservoir solution to 40%. Dehydrated crystals were mounted in nylon loops and flashcooled in liquid nitrogen. Data were collected at the Swiss Light Source X06SA beamline on a Mar225 CCD detector. Initial crystals displayed approximately 8 Å maximum resolution, but several lines of crystal improvement, such as dehydration and detergent mixtures, improved diffraction properties. Optimized crystals diffracted anisotropically to at least 3.3 Å in the best direction, and about 4.5 Å in the worst direction. Heavy-atom derivatives were obtained by adding $HoCl_3$, K_2PtCl_6 or Ta_6Br_{12} to the crystals before or during dehydration, either as salt or as a concentrated, aqueous solution.

Data processing. Data sets were processed using XDS[31]. The data quality was impaired by the strong anisotropy as also manifested by high R_{sym} values in the higher resolution bins (Supplementary Table 1). Initial heavy-atom positions were found using phases from a weak, low-resolution ($d > 8$ Å) molecular replacement solution using PHASER[32] and a partial search model derived from Ca^{2+}-ATPase[15]. Phases from the $HoCl_3$ and K_2PtCl_6 data were obtained by multiple isomorphous replacement with anomalous scattering (MIRAS). Phases from the Ta_6Br_{12} data were obtained by single isomorphous replacement with anomalous scattering (SIRAS). All phases was calculated by SHARP[33]. Several native data sets were used to yield optimal isomorphous pairing of individual derivative data sets. The heavy-atom-derived phases were refined, combined and extended at the maximum resolution of the native data by density modification using dmmulti[34], exploiting two-fold rotational non-crystallographic symmetry, a solvent content of 75% and several data sets displaying low levels of isomorphism for inter-crystal averaging. The resulting electron-density map was of high quality, providing a continuous trace of the main chain, albeit with a limited level of detail owing to the anisotropy of the data (Fig. 2, and Supplementary Fig. 4). Refinement was focused on the fitting of a model with reasonable stereochemistry to the experimental map. Before refinement the data were anisotropically corrected using the Anisotropy Correction Server[35] (http://www.doe-mbi.ucla.edu/~sawaya/anisoscale/). The model was built using O (ref. 36) with a Ca^{2+}-ATPase structure[15] and the CopA N domain structure[11] as guides for chain tracing. Initial torsion–angle refinement, imposing strict non-crystallographic symmetry, was performed in CNS1.2 (ref. 37) using only higher resolution reflections (5–3.6 Å) without bulk solvent correction. Iterative model building and refinement gradually improved the model and the fit to the experimental map. In later stages, bulk solvent correction was applied using phenix.refine[38] along with tight non-crystallographic symmetry restraints and use of the reflections in the 20–3.6 Å range. The final model yielded a crystallographic *R*-factor of 35.0% and a free *R*-factor of 36.6%. PROCHECK[39] evaluation of the ramachandran plot gave 52.8% in core regions, 38.0% in allowed regions, 8.1% in generously allowed regions and 1.1% in disallowed regions. Cavities in the model were located using Voidoo[40]. The full-length data were processed by XDS[31] and a molecular replacement solution was obtained using PHASER[32] and our model from the truncated form of AHA2. All figures were prepared using PyMOL[41].

31. Kabsch, W. Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J. Appl. Cryst.* **26,** 795–800 (1993).
32. Storoni, L. C., McCoy, A. J. & Read, R. J. Likelihood-enhanced fast rotation functions. *Acta Crystallogr. D* **60,** 432–438 (2004).
33. de La Fortelle, E. & Bricogne, G. Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. *Macromol. Crystallogr. A* **276,** 472–494 (1997).
34. Cowtan, K. 'dm': An automated procedure for phase improvement by density modification. *CCP4 ESF-EACBM Newsletter Prot. Crystallogr.* **31,** 34–38 (1994).
35. Strong, M. *et al.* Toward the structural genomics of complexes: Crystal structure of a PE/PPE protein complex from *Mycobacterium tuberculosis*. *Proc. Natl Acad. Sci. USA* **103,** 8060–8065 (2006).
36. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron-density maps and the location of errors in these models. *Acta Crystallogr. A* **47,** 110–119 (1991).
37. Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **54,** 905–921 (1998).
38. Afonine, P. V., Grosse-Kunstleve, R. W. & Adams, P. D. A robust bulk-solvent correction and anisotropic scaling procedure. *Acta Crystallogr. D* **61,** 850–855 (2005).
39. Laskowski, R. A., Macarthur, M. W., Moss, D. S. & Thornton, J. M. Procheck—a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26,** 283–291 (1993).
40. Kleywegt, G. J. & Jones, T. A. Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallogr. D* **50,** 178–185 (1994).
41. DeLano, W. L. The PyMOL molecular graphics system on the world wide web. ⟨http://www.pymol.org⟩ (2002).

Delivering the future

Drugs to treat diseases from cancer to AIDS could soon rely on short strands of RNA for their effects. But scientists must first work out how to navigate these fragments around the body. Nathan Blow reports.

The remarkable ability of short sequences of synthetic RNA to interfere with messenger RNA and thereby silence the activity of specific genes has proved incredibly helpful to geneticists wrestling with genetic function. And the push to harness this RNA interference (RNAi) for therapeutic use is now beginning to make headway. In the six years since the first paper reporting RNAi gene silencing in mammals was published[1], at least six therapeutic programmes based on the concept have moved into clinical trials.

"Progress in the field of RNAi therapeutics has occurred remarkably fast," says John Maraganore, president and chief executive of Alnylam Pharmaceuticals in Cambridge, Massachusetts. But delivering the sequences remains a problem. Initial clinical trials relied on 'local delivery', directly introducing short interfering RNAs (siRNAs) into the specific tissue they were to treat. But for true therapeutic value, the siRNAs need to be introduced systemically.

"Systemic delivery is the major issue right now," says Alan Sachs, vice-president for RNA therapeutics based at Sirna Therapeutics, a wholly owned subsidiary of Merck in San Francisco.

Getting a small RNA to interfere with the right messenger RNA in the correct tissue and cell type at a safe, therapeutic level by systemic administration requires an exquisite degree of control — creating the need for different delivery vehicles and potentially even specialized targeting strategies. Animal studies[2] have shown that it is possible for siRNAs delivered systemically to silence target genes. "What we have learned over the past couple of years is that systemic delivery of RNAi can be achieved, and there are a variety of methods that can be used to achieve it," says Maraganore. But he is also quick to note that there is no simple solution.

"If you inject naked siRNA into the blood, under normal pressure, it doesn't work," says Daniel Anderson of the Center for Cancer Research at the Massachusetts Institute of Technology (MIT) in Cambridge. Yet naked siRNA delivered directly into the lungs to treat respiratory syncytial virus (RSV) can profoundly reduce viral replication. This contrast highlights the chasm between local and systemic delivery of synthetic siRNAs.

MERCK

Alan Sachs says that delivery is the major issue for RNAi therapeutics.

Local delivery

Alnylam focused on local delivery when it began to develop RNA-based therapeutics. Its treatment for RSV in the lungs uses an inhaled synthetic RNA that triggers the destruction of a protein essential for the virus's replication. Inhalation allows high local concentrations of the RNA to be achieved, says Maraganore, while also taking advantage of naturally occurring mechanisms, such as pinocytosis, for uptake into the target cells.

Work by researchers at the company has also helped shed light on systemic delivery. In a 2004 study[2], they affirmed for the first time the potential 'drug-like' properties of siRNAs when

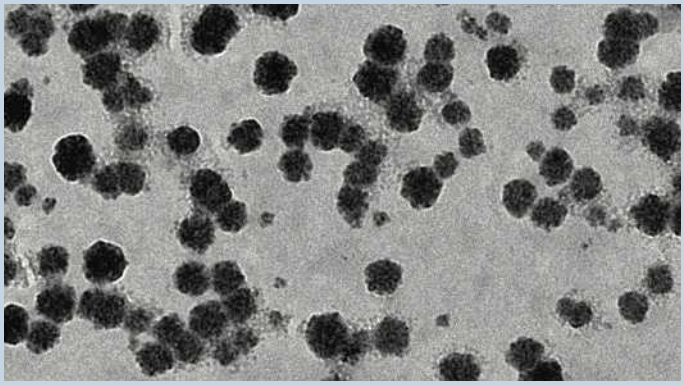
THE VEHICLE LABORATORY

Many companies are realizing that the development of new delivery vehicles for therapeutics based on RNA interference (RNAi) will require collaborative efforts. "We view the delivery of small RNAs as one of the most important biomedical endeavours in modern biological science," says John Maraganore, chief executive of Alnylam Pharmaceuticals in Cambridge, Massachusetts. "And when you have such a broad-scale challenge and opportunity, you can't do everything internally — you have to work with the best groups outside as well."

This spring, Alnylam initiated a collaboration with Daniel Anderson and Robert Langer at the Massachusetts Institute of Technology (MIT) in Cambridge. Under the agreement, Alnylam will provide funding for ten postdocs at MIT to work on issues directly related to small-RNA delivery while having the exclusive option to any intellectual property that comes out of the effort.

Anderson says that at its core this is an academic programme, so the goal is to get postdocs to work on the delivery issue while being trained as future leaders in the field.

"One nice feature is that we have collaborations with a company that is a leader in this field, so it allows us to accelerate our efforts tremendously," Anderson says. For example, when the MIT researchers develop any new delivery technologies, scientists at Alnylam can provide animal models and other methods to test and evaluate these new vehicles, providing quick feedback on whether or not the research is on the right track.



NANO LETT. 7, 874–879 (2007).

An electron microscope image of nanoparticles that can be used as delivery vehicles for siRNA.

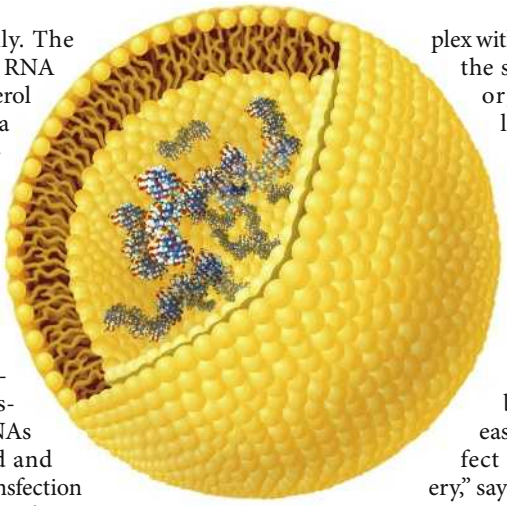
Langer's lab has a long history of translational research. It has developed principles that have led to some 40 products that are now in clinical trials or have been approved by the US Food and Drug Administration. "The hope is that we can do that here," says Langer. "We can do the kind of basic research to help solve the RNAi delivery problem and then work closely with Alnylam to take the basic research into the clinic where it can be used to treat different diseases."

In addition to the MIT collaboration, Alnylam also funds R&D and manufacturing activities in Vancouver, Canada, to further the development of cationic lipids for delivery, and has some 25 feasibility agreements in place in which Alnylam is testing and evaluating technology that has been introduced from either companies or academic groups. N.B.

delivered systemically. The team used a synthetic RNA conjugated to cholesterol and stabilized with a partial phosphorothioate backbone and 2′-*O*-methyl sugar modifications on both the sense and antisense strands of the RNA.

Since this study, both lipid- and polymer-based vehicles for systemic delivery of siRNAs have been developed and tested. At Polyplus-transfection in Illkirch, France, researchers have taken advantage of the difference between cationic polymers and cationic lipids for systemic delivery to different organs. "We are interested in delivery to the lung and have used systemic administration of the cationic polymer polyethylenimine for delivery," says Patrick Erbacher, the company's chief scientific officer. "But for tumour injections, we use either a cationic polymer or a cationic lipid formulation."

With siRNAs conjugated to lipids or encapsulated in liposomes or lipid nanoparticles, several companies have achieved stable and efficient systemic delivery to organs including the liver, pancreas, kidneys and even to some types of tumour. And polymers that can complex with siRNA can deliver the short sequences to organs such as the lungs, spleen and kidneys.



Liposomes offer one way of achieving systemic delivery of siRNAs.

ENCAPSULA NANOSCIENCES

Researchers at Altogen Biosystems based in Las Vegas, Nevada, are exploring cationic lipids and biodegradable polymers for *in vivo* delivery, but have not found it easy. "There is no perfect method for delivery," says Andreas Kim, the company's vice-president of research and development.

"Nothing really works amazingly well. All methods have their advantages and disadvantages." In mice, he notes, lipid-based delivery of siRNA is very efficient but tends to induce an inflammatory response to the lipid formulation. Delivery vehicles based on biodegradable polymers, on the other hand, don't cause inflammatory responses but are not delivered as efficiently and the effects seem to be more transient than their lipid-based counterparts.

Alnylam is using liposomes to deliver siRNAs to the liver. "It is easier to target things in the liver with liposomes because about 95% of the injected dose for liposomal formulations goes to the liver," Maraganore says. Liposomes are synthetic analogues of the cell membrane and are made up of hydrophilic and hydrophobic regions that form spherical 'packages' in aqueous conditions. Alnylam is using this approach to target two genes in the liver — one involved in regulating levels of low-density lipoprotein in the blood and the other involved in liver cancer (targeting both vascular endothelial growth factor and kinesin spindle protein).

The ability to use lipid-based delivery vehicles to target the liver has made the organ a popular starting point for many companies. Merck, for example, is using lipid nanoparticles that, like liposomes, take advantage of endogenous mechanisms for uptake, but have no specific ligand attached for targeting.

John Maraganore is confident systemic delivery of siRNA can be achieved.

ALNYLAM PHARMACEUTICALS

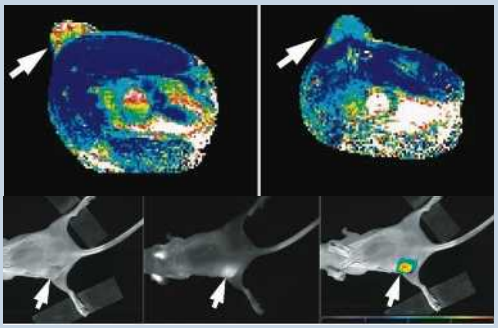
Although lipid nanoparticles are its main focus, Merck is also exploring other delivery vehicles — in some cases through external collaborations. "We are working aggressively in the licensing arena — inviting people to work with us in an evaluation phase," says Sachs. In October, the company finalized a licensing agreement with

THE INSIDE TRACK

How to deliver short-interfering RNAs (siRNAs) to specific tissues is only part of the problem facing researchers. They also need to find out whether the RNA has reached its intended target. Anna Moore, a radiologist at Harvard Medical School in Boston, Massachusetts, is well aware of the issue. "There was no way to use a clinical imaging modality to see the delivery of siRNA," she says.

When researchers want to see whether an siRNA has been reached a particular tissue, they usually perform histological analysis followed by reverse transcription PCR to see whether the target gene was silenced. "You can do this with mice, but when you move on to humans it becomes impractical," Moore says.

Researchers can track siRNAs *in vivo* using bioluminescence imaging or by tracking green fluorescent proteins. But bioluminescence imaging is not a clinical modality. So Moore and her colleagues decided to try magnetic resonance imaging (MRI).



Now you see it: the nanoparticle system devised by Anna Moore's team allows siRNA delivery to be seen in MRI scans (top) and optical scans (bottom).

NATURE MED. (REF. 5)

The first step was to design an siRNA delivery vehicle that could be imaged by MRI. Moore and her team used a nanoparticle containing an iron oxide core. They coated it with dextran, which could have various targeting features added to it relatively easily.

Although iron oxide can be imaged using MRI, the group also attached a fluorescent dye, Cy 5.5, to the dextran coat for optical imaging. "We wanted to correlate the imaging data with microscopic findings," says Moore.

The iron oxide nanoparticle generates a bright spot on the MRI image. The exact target of the nanoparticle can then be confirmed by the fluorescent dye and by doing microscopy for histological analysis.

Two further attachments were then made to the nanoparticle via the dextran coating: a membrane translocation peptide that can cross cell membranes and an siRNA. With this, Moore and her colleagues thought they had a particle that could target and image delivery to tumour cells *in vivo*. But the imaging showed that the nanoparticle went to the liver and kidneys, and was present in other organs as well[5].

Moore and her team plan to continue with the nanoparticles, trying to make them more efficient in terms of delivery and target uptake. But the real value of these nanoparticles might be their versatility. As different siRNAs or targeting peptides can be attached to the dextran coat, a large range of therapeutic siRNAs and peptides can be tested.

"My lab is really interested in imaging other pathologies such as diabetes, which is far from cancer but the imaging approaches are very similar," says Moore. "And that is the beauty of this technology — you can apply it to different pathologies." N.B.

Protiva Biotherapeutics of Burnaby, British Columbia in Canada, for its stable nucleic acid lipid particles (SNALPs). These are specialized lipid nanoparticles that encapsulate siRNA and were the first non-viral siRNA delivery vehicles that showed activity in non-human primates[3]. Alnylam has also launched a number of academic and industrial delivery collaborations (see 'The vehicle laboratory').

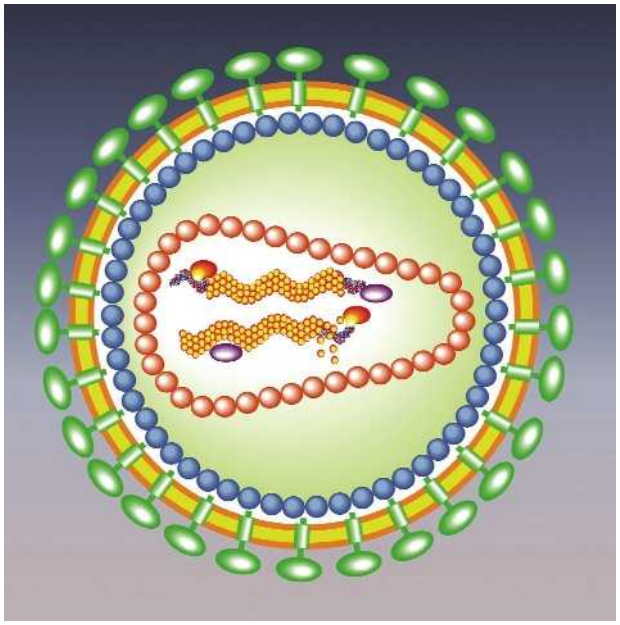
The rapid advancement of RNAi-based therapeutics is leading Polyplus-transfections to explore the manufacturing aspects of delivery vehicles. The company develops and markets DNA, RNA and protein transfection and delivery reagents for both *in vitro* and *in vivo* applications. And with RNAi therapeutics on the cusp of entering the clinic, the company sees the need to produce delivery vehicles in bulk under governmental quality specifications or 'current good manufacturing practices' (cGMP) standards. "If you want to be able to get these into the clinics, you have to have cGMP-qualified delivery systems," says Erbacher.

Beyond the liver

Despite this progress in general delivery, there is still some way to go for full therapeutic application. "I do not think that we are at the point now where we can name a tissue, specifically deliver and get good knockdown," says Anderson. Although he notes that there is good evidence to suggest that targeted systemic delivery should be possible.

Kim agrees that targeted delivery looks promising, but he cautions that the development of the delivery vehicles will be complicated. "You need a number of additional steps in the formulation process," he says.

A promising approach to targeted delivery involves incorporating targeting elements such as antibody fragments, ligands and small chemical groups with the synthetic RNA. At the same time, techniques are being designed to image targeted nanoparticle delivery and biodistribution (see 'The inside track'). Some of the new approaches might overcome issues associated with lipid-based vehicles. For example, RNA aptamers attached to siRNAs, which have been used by researchers to target prostate cancer cells, might not cause the inflammatory responses seen in some cases with lipids or antibodies.



I. VERMA

Lentiviruses, shown here schematically, can be modified to carry RNA into cells for potential therapeutic effect.

Another approach that might avoid inflammatory responses is being developed by Calando Pharmaceuticals in Pasadena, California, using polymers that contain cyclodextrin. When the polymers are mixed with siRNAs, they bind to the RNA backbone and assemble into nanoparticles. Targeting ligands or even stabilizing agents can then be attached to the cyclodextrin in the polymer to improve delivery.

The evolutionary advantage

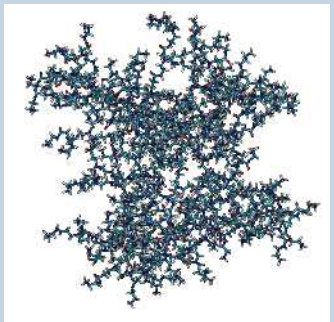
Another approach to targeted delivery takes advantage of nature. "Viruses have learned how to get into cells — cross the membrane into the cytoplasm then get into the nucleus," says Inder Verma, a geneticist at the Salk Institute for Biological Studies in La Jolla, California. "All other approaches have to figure out how to do this." This gives viruses an "evo-

THINKING SMALL

"Each day you see major, significant publications on the roles of microRNAs in regulating pathways of genes involved in disease aetiology," says John Maraganore of Alnylam Pharmaceuticals in Cambridge, Massachusetts. Alongside the research into the mechanisms and roles of microRNA (miRNA) in disease, researchers are now starting to look at the potential of miRNA-based therapeutics.

Regulus Therapeutics in Carlsbad, California, is one of the first companies to be founded entirely for the development of miRNA-based therapeutics. It was formed in September as a joint venture between Alnylam and Isis Pharmaceuticals in Carlsbad. "We recognized that miRNAs were becoming a potential therapeutic opportunity," says Frank Bennett, senior vice-president of research at Isis, "but we also recognized that it was difficult for both companies to input the resources that were warranted." After much discussion, the two firms agreed to supply miRNA assets and core technologies to Regulus.

Regulus is taking two approaches to the development of miRNA-based therapeutics. "The most advanced approach is inhibiting the function of an endogenous miRNA in cells," says Bennett. This uses a synthetic oligonucleotide to target the miRNA for silencing.



C. KELLY, B. ORR & M. B. HOLL

Testing of technologies, such as dendrimers, for delivering siRNAs may pave the way for miRNA drugs.

Although this is different from siRNA, as the miRNA is the target not the therapeutic agent, Bennett thinks that much of the technology developed for targeting mRNAs with siRNA is directly translatable to the targeting of miRNAs with oligonucleotides. And he notes that using oligonucleotides will also minimize any 'off-target' effects. "These oligonucleotides are very specific — even a single base mismatch will cause the oligo to lose activity," he says.

The second approach involves replacing miRNAs in or delivering them to cells, which Bennett says is similar to the current approach for siRNAs. The first applications here might be replacing miRNAs that are missing from disease-associated cells but present in a normal cells, which happens in some cancers, or augmenting naturally occurring miRNAs.

Although experience of siRNAs will help this nascent field, there is one unique issue in miRNA biology that researchers will have to address: miRNAs can inhibit tens to hundreds of genes at a time. Scores of researchers, both inside and outside academia, are now working to understand exactly how this regulation by miRNAs occurs. "You can use specific chemical modifications to the miRNA to limit that potential," says Bennett, "but it is still present."

Many researchers and companies now think miRNA-based therapeutics hold great promise for the future. And although some companies are waiting in the wings for more information on their basic mechanisms of action, others including Merck, Santaris Pharma of Hørsholm, Denmark, Rosetta Genomics in Rehovot, Israel, and Actigenics of Maurens-Scopont, France, are actively involved in miRNA research and development. N.B.

lutionary advantage" as delivery vehicles for small RNAs, he adds.

By exploiting the natural diversity of viruses, researchers have developed a suite of viral vectors that can be used to deliver therapeutic sequences into a wide variety of cell types. Viral vectors have been around for some time and scientists are currently using them to deliver short hairpin RNA (shRNA) that can be processed by cells to yield siRNAs and silence genes.

Viral vectors from the DNA-based adenovirus and adeno-associated virus (AAV) have been used by researchers and companies, such as ArmaGen Technologies in Santa Monica, California, for shRNA delivery. AAV vectors have, however, found more widespread use in RNAi applications as these viruses can transduce both dividing and non-dividing cell types and result in stable, site-specific integration.

Lentivirus-based vectors are another option. "The utility of lentiviral vectors for RNAi research is their ability to transduce non-dividing cells and to efficiently transduce difficult-to-transfect cells," says Boro Dropulic, founder and chief executive of Lentigen in Baltimore, Maryland.

But any viral vector will still have to overcome the issue that plagues synthetic RNAs: targeting. "Targeting is a big problem. There are several groups working on that issue now but the jury is still out," says Dropulic. What makes targeting viral vectors such a challenge is the difficulty in altering the viral structure with targeting elements while maintaining the proper function of viral particles.

For adenovirus and AAV vectors, for example, the viruses require a specific number of proteins to make their viral shell. Adding more proteins, such as an antibody fragment for targeting purposes, could be problematic. "Imagine the viral shell is like the dome of a sports centre," says Verma. "You can only put so much information in there, after which the dome will break." To overcome this problem, Verma thinks that researchers will need to find a way to modify the entire shell of the virus.

Lentiviral vectors are less problematic — the virus particles are surrounded by a lipid envelope in which proteins can be inserted without disrupting the viral structure. Even so, efficient and specific targeting remains an issue.

I. VERMA

Inder Verma believes viruses have an evolutionary advantage for RNAi delivery.

MERCK

Sirna Therapeutics is developing stable siRNA compounds for silencing disease-related genes.

"I think that it will come down to two things: a more extensive learning of the modification of the viral structure proteins, and identification of specific target receptors so that you can have the viruses interact with cells in a very specific manner," says Verma. He points to recent work from David Baltimore's lab at the California Institute of Technology in Pasadena, in which a small monoclonal antibody was linked to a lentiviral envelope protein for targeting to specific cell types[4], and to other groups trying to use small ligands and signature sequences from cell-surface receptors as promising developments for the targeting of viral vectors.

As researchers wrestle with the targeting issue, some companies have already begun to use lentiviral vectors to deliver therapeutic RNAi. Dropulic sees the *ex vivo* use of lentiviral transduction systems as being a critical first step. "I think using the transduced cells as the vehicles of widespread dissemination rather than thinking about direct injection of lentiviral vectors will be the way to go," he says. Lentigen is using this approach on T cells and stem cells for cancer therapies as well as for treatment of infectious diseases.

The silent future

Most experts agree that therapeutic RNAi will probably not rely on a single delivery vehicle or administration approach for all diseases. "Some diseases might require lower doses than others or could be semi-local in nature," says Robert Langer, a bioengineer at MIT, "so I think a lot may depend on the disease and treatment modality you are thinking about."

One likely beneficiary of the work on delivery mechanisms are siRNA's recently uncovered siblings: microRNAs (miRNAs), which are naturally occurring siRNAs found in plants and animals. "With miRNA therapeutics, you are generally focusing on a new class of non-coding RNAs where the biology is still being discovered," says Maraganore. It is early days for this technology, but companies are now being established to work exclusively on miRNA-based therapies (see 'Thinking small').

When it comes to siRNA, Maraganore is encouraged by the early approaches to delivery being used in clinical studies. He expects the development of new delivery vehicles to continue for sometime to come. "I suspect that in 35 years scientists will continue to work on optimizing delivery approaches," he says.

And Verma thinks that the discovery of RNAi has even breathed fresh life into the still struggling field of gene therapy. "I think having RNAi technology available now gives a new impetus because it is such an effective technology — that is, if only we could deliver it." ■

Nathan Blow is technology editor for *Nature* and *Nature Methods*.

1. Elbashir, S. M. *et al. Nature* **411,** 494–498 (2001).
2. Soutschek, J. *et al. Nature* **432,** 173–178 (2004).
3. Zimmermann, T. S. *et al. Nature* **441,** 111–114 (2006).
4. Yang L., Bailey, L., Baltimore, D. & Wang, P. *Proc. Natl Acad. Sci. USA* **103,** 11479–11484 (2006).
5. Medarova, Z., Pham, W., Farrar, C., Petkova, V. & Moore, A. *Nature Med.* **13,** 372–377 (2007).

Correction

In the Technology Feature 'The personal side of genomics' (*Nature* **449,** 627–630; 2007) we said that Illumina uses emulsion PCR in its next-generation sequencing systems. In fact, the company's genome analyser uses solid-phase amplification on a planar, optically transparent surface.

COMPANY	PRODUCTS/ACTIVITY	LOCATION	URL	
RNAi therapeutic development				
Actigenics	Discovery and evaluation of miRNAs that alter gene expression for use in therapeutic development	Maurens-Scopont, France	www.actigenics.com	
Acuity Pharmaceuticals	RNAi-based treatments for age-related macular degeneration and diabetic retinopathy	Philadelphia, Pennsylvania	www.acuitypharma.com	
Alnylam Pharmaceuticals	siRNA–based therapeutics for RSV, liver cancer and hypercholesterolaemia	Cambridge, Massachusetts	www.alnylam.com	
ArmaGen Technologies	Developing technologies for siRNA delivery across the blood–brain barrier	Santa Monica, California	www.armagen.com	
Benitec	RNAi-based therapeutics for infectious disease, neurological disorders, cancer and autoimmune diseases	Melbourne, Australia	www.benitec.com	
Calando Pharmaceuticals	Technologies for the therapeutic use of RNAi	Pasadena, California	www.calandopharma.com	
Cequent Pharmaceuticals	Technologies for targeted delivery of RNAi using non-pathogenic bacteria to produce interfering RNA	Cambridge, Massachusetts	www.cequentpharma.com	
Intradigm	Systematic RNAi-based therapeutics focused on oncology	Rockville, Maryland	www.intradigm.com	
Isis Pharmaceuticals	RNA-inhibiting therapeutics; Vitravene, the first antisense drug to receive market clearance	Carlsbad, California	www.isispharm.com	
Merck	RNAi therapeutics using lipid-nanoparticle delivery	Rahway, New Jersey	www.merck.com	
NeoPharm	Cancer drugs and therapeutics using liposome-based delivery	Lake Forrest, Illinois	www.neophrm.com	
Regulus Therapeutics	miRNA-based therapeutics	Carlsbad, California	www.regulusrx.com	
Rosetta Genomics	miRNA therapeutics and diagnostics related to cancer	Rehovot, Israel	www.rosettagenomics.com	
RXi Pharmaceuticals	RNAi-based therapeutics to treat human diseases with an initial focus on neurodegenerative disease, oncology, diabetes and obesity	Worcester, Massachusetts	www.rxipharma.com	
Santaris Pharma	Assays for disease-related miRNAs as targets; designing and making miRNA antagonists	Copenhagen, Denmark	www.santaris.com	
Silence Therapeutics	Novel siRNA called AtuRNAi; therapeutic RNAi programme concentrated on cancer	Berlin, Germany	www.silence-therapeutics.com	
VirxSys	Therapeutics using lentiviral vectors for gene-therapy applications	Gaithersburg, Maryland	www.virxsys.com	
Tools and services for RNAi research				
Ambion	Silencer range of kits, vectors, siRNAs and reagents for RNAi; kits and products for RNA synthesis, isolation, quantitation and analysis	Austin, Texas	www.ambion.com	
Amaxa	Transfection reagents, siRNA test kits, culture media	Gaithersburg, Maryland	www.amaxa.com	
Applied Biosystems	Pre-designed and validated siRNAs; miRNA-expression reporter vector system; real-time PCR miRNA assays	Foster City, California	www.appliedbiosystems.com	
Asuragen	RNA-based therapeutic and diagnostics company with a core focus on miRNA	Austin, Texas	www.asuragen.com	●
BD Biosciences	Culture media, FACS range of flow cytometers; RNAi kits	San Jose, California	www.bdbiosciences.com	
Cenix BioScience	High-throughput applications of RNAi; siRNA design	Dresden, Germany	www.cenix-bioscience.com	
CombiMatrix	miRNA microarrays from various species	Mukilteo, Washington	www.combimatrix.com	
Dharmacon	Smartvector shRNA lentiviral particles for delivery; RNAi screening services; reagents for miRNA functional analysis	Lafayette, Colorado	www.dharmacon.com	
Eurogentec	Kits for amplification of human pre-miRNA; custom siRNA	Seraing, Belgium	www.eurogentec.com	
Exiqon	Gene-expression analysis services; identification of new miRNAs	Vedbaek, Denmark	www.exiqon.com	●
GeneTools	Morpholino antisense oligonucleotides for blocking miRNA	Philomath, Oregon	www.gene-tools.com	
genOway	Services for gene knockdown in mice using RNAi technologies	Lyon, France	www.genoway.com	
Imgenex	Plasmid-based RNAi-expression vectors, prepared adenoviral vectors for shRNA delivery	San Diego, California	www.imgenex.com	
Integrated DNA Technologies	Small RNA cloning kit, antisense oligonucleotides	Coralville, Iowa	www.idtdna.com	
Invitrogen	Synthetic RNA, RNAi vectors, RNAi-knockdown services	Carlsbad, California	www.invitrogen.com	
Invivogen	psiRNA system for plasmid-based delivery of siRNA in mammalian cells	San Diego, California	www.invivogen.com	
Lentigen	Development of lentiviral vectors for RNA delivery	Baltimore, Maryland	www.lentigen.com	
Mirus Bio	RNAi-knockdown services in animals; products for nucleic-acid isolation, gene transfer and RNAi	Madison, Wisconsin	www.genetransfer.com	●
MWG Biotech	Custom and pre-designed siRNA	Ebersberg, Germany	www.mwg-biotech.com	
New England Biolabs	Kits for the generation of RNA for RNAi experiments	Beverly, Massachusetts	www.neb.com	●
OligoEngine	pSUPER RNAi vector system for gene silencing; custom oligonucleotides for all applications	Seattle, Washington	www.oligoengine.com	
Open Biosystems	RNAi libraries; lentiviral-packaging systems for shRNA delivery	Huntsville, Alabama	www.openbiosystems.com	
Polyplus-transfection	Transfection reagents for DNA, RNA and protein; developing siRNA delivery vehicles using cationic lipids and polymers	Illkirch, France	www.polyplus-transfection.com	
Promega	RNAi vectors and expression systems	Madison, Wisconsin	www.promega.com	
QIAGEN	Kits for the isolation, purification and labelling of miRNAs; pre-and custom-designed siRNAs; siRNA-transfection reagents	Germantown, Maryland	www1.qiagen.com	

COMPANY	PRODUCTS/ACTIVITY	LOCATION	URL	
Sigma Genosys	Custom siRNA synthesis; siRNA design service	The Woodlands, Texas	www.sigmaaldrich.com	●
Stratagene	Tools and reagents for molecular-biology research; siRNA transfection system	La Jolla, California	www.stratagene.com	●
Sylentis	SIRFINDER technology to determine siRNAs with potential pharmacological applications	Madrid, Spain	www.sylentis.com	
Targeting Systems	Targefect siRNA transfection system	Santee, California	www.targetingsystems.com	
Vectalys	Viral-vector production and cell transduction	Labege, France	www.vectalys.com	

General

Agilent	Instrumentation for genomics and proteomics research	Santa Clara, California	www.agilent.com	
Beckman Coulter	Tools and systems for molecular biology, genomics and proteomics research	Fullerton, California	www.beckmancoulter.com	
Bio-Rad	Products, instruments and software for life-sciences research	Hercules, California	www.bio-rad.com	●
BMG Labtechnologies	Microplate and array readers and handling systems	Offenburg, Germany	www.bmglabtech.com	
Brinkmann Instruments	Laboratory-instrument supplier; consumables	Westbury, New York	www.brinkmann.com	
Carl Zeiss	Imaging systems	Jena, Germany	www.zeiss.com	
EMD Biosciences	Calbiochem, Novabiochem, and Novagen product lines	San Diego, California	www.emdbiosciences.com	
Epicentre Biotechnologies	Enzymes for PCR and RT-PCR; DNA and RNA purification	Madison, Wisconsin	www.epibio.com	
Eppendorf	Consumables for molecular biology; instrumentation	Hamburg, Germany	www.eppendorf.com	
Gilson	Pipettes, automated liquid handling, liquid-chromatography systems and software	Middleton, Wisconsin	www.gilson.com	
Hamilton Company	Automated liquid-handling stations	Reno, Nevada	www.hamiltoncompany.com	
Millipore	Reagents and kits for cell biology, genomics and immunodetection	Billerica, Massachusetts	www.millipore.com	●
MP Biomedicals	Reagents and chemicals for research	Aurora, Ohio	www.mpbio.com	●
New Brunswick Scientific	Cell-cultivation systems, bioreactors, incubators, freezers	Edison, New Jersey	www.nbsc.com	
Nikon Instruments	Microscopy instrumentation, accessories and analysis software	Melville, New York	www.nikoninstruments.com	●
PerkinElmer Life Sciences	Instruments, reagents and kits for life-sciences research	Waltham, Massachusetts	las.perkinelmer.com	
Pierce Chemical	Protein assays, purification, Western blotting	Rockford, Illinois	www.piercenet.com	
Roche Diagnostics	Reagents and kits for molecular biology; genomics instrumentations and software	Lewes, UK	www.roche-applied-science.com	
Sigma Aldrich	Reagents and kits for molecular-biology research; chemicals for life-sciences research	St Louis, Missouri	www.sigmaaldrich.com	●
Takara Bio	Reagents, kits and consumables for molecular biology	Shiga, Japan	www.takara-bio.com	
Thermo Fisher Scientific	Life-sciences research consumables, automation and robotics, chemicals and consumables	Waltham, Massachusetts	www.thermofisher.com	
USB	Chemicals and reagents for molecular-biology research	Cleveland, Ohio	www.usbweb.com	
Wako Chemicals	Speciality chemicals supplier; clinical diagnostic reagents	Richmond, Virginia	www.wakousa.com	

● see advertisement

naturejobs

JOBS OF THE WEEK

"It's *your* degree," Nicole Christacos told the audience of postdocs. "Do with it what you want." Christacos, a cytogeneticist at Quest Diagnostics in Madison, New Jersey, was speaking at a postdoc careers symposium held last week in Washington DC by the American Society for Cell Biology, and she emphasized this mantra more than once. It's one that makes perfect sense.

The panellists at the meeting shared their success stories, describing how they used their degrees to get the jobs they wanted. Christacos went into industry despite a lack of encouragement from her peers. Kavita Berger, who has a PhD in medical genetics, was determined to get into science policy. After unsuccessfully seeking jobs in government, she was ready to settle for an internship at the American Association for the Advancement of Science, but was told she was overqualified. Fortunately, another position opened up in the association's Center for Science, Technology and Security Policy. Her move to Washington DC and networking with others in policy has helped her to further her career.

And John LeGuyader became a patent examiner at the United States Patent and Trademark Office (PTO) after leaving academia and deciding high-school teaching wasn't for him. He thought he might go to law school, but a mentor encouraged him to approach the PTO first, as the office often pays for employees to attend law school. Seventeen years later, he has had a stimulating and stable career at the PTO — although he never made it to law school. LeGuyader used his training and technical know-how to make a career determining what is patentable and why.

No matter how many times graduate students hear about worsening job prospects in academia and the importance of alternative career paths, the regimented structure for advancement in academic science tends to make young scientists defer to sometimes short-sighted mentors on their career direction. This could close off avenues to industry or other 'alternative' careers. But as the panellists made clear, if you maintain your awareness of alternative paths and remain steadfast, you can use your degree to get the career you want.

Gene Russo, acting editor of *Naturejobs*

CONTACTS
Acting Editor: Gene Russo

European Head Office, London
The Macmillan Building,
4 Crinan Street,
London N1 9XW, UK
Tel: +44 (0) 20 7843 4961
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

European Sales Manager:
Andy Douglas (4975)
e-mail: a.douglas@nature.com
Business Development Manager:
Amelie Pequignot (4974)
e-mail: a.pequignot@nature.com
***Nature*events:**
Claudia Paulsen Young
(+44 (0) 20 7014 4015)
e-mail: c.paulsenyoung@nature.com

France/Switzerland/Belgium:
Muriel Lestringuez (4994)
Southwest UK/RoW:
Nils Moeller (4953)
Scandinavia/Spain/Portugal/Italy:
Evelina Rubio-Hakansson (4973)
Northeast UK/Ireland:
Matthew Ward (+44 (0) 20 7014 4059)
North Germany/The Netherlands:
Reya Silao (4970)
South Germany/Austria:
Hildi Rowland (+44 (0) 20 7014 4084)

Advertising Production Manager:
Stephen Russell
To send materials use London address above.
Tel: +44 (0) 20 7843 4816
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com
***Naturejobs* web development:**
Tom Hancock

***Naturejobs* online production:**
Jasmine Myer
US Head Office, New York
75 Varick Street, 9th Floor,
New York, NY 10013-1917
Tel: +1 800 989 7718
Fax: +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

Japan Head Office, Tokyo
Chiyoda Building,
2-37 Ichigayatamachi,
Shinjuku-ku, Tokyo 162-0843
Tel: +81 3 3267 8751
Fax: +81 3 3267 8746

Asia-Pacific Sales Manager:
Ayako Watanabe
Tel: +81-3-3267-8765
e-mail: a.watanabe@natureasia.com

Up, down, and out

As my period of writing postdoc journals draws to a close, I wonder whether readers who followed my entries through the year would have been surprised by my final in-print journal entry (see *Nature* **450,** 582; 2007). In that entry, I disclosed my decision to leave science research, a decision I made with much trepidation.

Looking back at my journal entries, I do detect hints — some obscure, some less so — of my growing discontentment with science research. Sometimes I told the reader a story that portrayed my situation as more encouraging than it actually was. I didn't want to reveal my restlessness in such a public forum, especially as I hadn't been in the job very long. Had I thrown prudence to the wind, and had I had more space to elaborate, the surprise ending of my narrative wouldn't have been much of a surprise. Rather, readers would have seen that my story was akin to that of a lone hiker in desert country who scales the occasional sublime peak, but who more often than not endures the tedium of a featureless and desolate landscape.

Don't misunderstand me: the peaks have presented some quite spectacular views. During the past year I have had the distinct pleasure of working with people who are wholly committed to their research, people who would do practically anything through their work to help eradicate the health scourges afflicting our society. Mixing their passion and intelligence with a strong dose of luck, I have been fortunate enough to have a manuscript describing part of my research accepted for publication in a reputable journal. With a little more good fortune, I'll submit at least one more paper before I leave the lab. Although I won't be around long enough to see whether my published articles will effect change — a criterion I listed in my very first journal entry for determining whether I am making a palpable difference in the world — I'm quietly hoping that at least one other person in the research community will find my publications of some use.

No, it's not the exhilarating peaks that are the problem. In fact, it's not even the tedium of the unremarkable valleys — after all, every job, no matter how wonderful, has its less satisfying aspects. I've simply come to the conclusion that I'd rather not be hiking along this particular path at all.

There are at least two reasons why. First, for me, basic research is too disconnected from the daily lives and experiences of the people whose diseases and afflictions we seek to understand. The less connected my work is to the immediate needs and concerns of people, the stronger my urge to become *more* connected. Finding a line of work that allows me to regularly interact with the people whose lives I'm trying to change is a major priority.

Second, no matter how exciting those research breakthroughs may be, my dissatisfaction with life in the lab has been accompanied by a growing interest in fields of endeavour outside the purview of science. My narrow high-school and university education in engineering and the natural sciences meant that I was only superficially exposed to entire domains of thought and practice — philosophy, anthropology, politics, history, literature and religion, among many others — whose importance for living wisely and participating whole-heartedly in a democratic society I am only now beginning to appreciate. My curiosity about the natural world remains strong, but I also have other interests.

And so I bid farewell to the world of science research, at least for now. It was great fun while it lasted.

Peter Jordan is a visiting fellow at the National Institute of Diabetes and Digestive and Kidney Diseases in Bethesda, Maryland.

Recoper

Breathing life into the revolution.

Neal Asher

When the stealth boat rose on its hydrofoils, the wind and spray kept me cool in the bright African sun. I gazed back and saw that the Eugov gunboat had finally given up the chase.

Jansen grinned at me. "We're in Moroccan territory now."

Memtech initiated the first recoper in 2044, the year the National Health Police seized a 1,000-tonne shipment of Argentinian beefburgers and subsequently smashed the notorious Midlands fried-food ring, which was led, as government-approved blogs delighted in telling us, by the 'Yorkshire Chipper'. At this time my wife, Gillian, announced the happy news that CCTV would be installed in our flat — she worked for CPHS (Camera Partnership for Home Safety) and had volunteered our place as a test bed.

The recoper was Mohammed Aswar MacDoogal and, as I wrote his biography on Wikibio, Memtech, never revealing their true purpose, paid Eugov for my expertise. Like every European citizen I was a state employee but, being leased to a private company and actually generating wealth, I was also a 'societal asset', which meant filing notice of all my movements and work-related activities a week beforehand. This was heartbreaking, as I'd been about to suggest to Gillian that we escape to North Africa on one of the refugee boats. It never occurred to me that there might be a connection between my work and the CPHS cameras in our flat.

MacDoogal was a notorious libertarian blogger whose attacks on the formation of Eugov caused much chagrin in Notting Hill champagne and socialism circles. He was born to a Calvinist Scottish father and an Islamic Pakistani mother and in public claimed to be a Sikh — although privately he admitted this was so he could carry a dagger and didn't have to wear a crash helmet when thrashing his 1,000 cc antique Ducati motorbike about the Highlands. He started his blog 'Invisible Worm' in 2008 with an article dissecting the then €1.2-billion cost of the British Olympics. Over the ensuing 20 years he wrote more than 8 million words, created numerous animations, short films and video news reports, in all of which he never revealed his identity. His blog is huge, and even now I have not seen all of it, for its thousands of distracting hyperlinks make this a near-impossible task.

Working for Memtech I became hugely frustrated by the byzantine Diversity and Equality regulation, which had become suffocating after I wrote MacDoogal's biography. But Memtech, which we now know was a front for American-financed revolutionary group Free Europe, wanted the truth about MacDoogal, and risked telling me their true aim.

JACEY

I loved the idea and obliged them by first providing the insipid and politically correct version, which I transmitted via e-mail, next providing the real deal, which I put on a memchip and took directly to their office in Hastings. Foolishly, I told Gillian about this subterfuge and, on a subsequent visit to Memtech, Jansen apprised me of the reality.

"Once we've got all we need we'll run the recoper and transmit it all out-state, and MacDoogal will soon be a thorn in Eugov's side again," he said. "Then, of course, we'll have to get out."

"I do have a wife," I told him.

"Yes," he said, "the one who had Home Safety CCTV installed to keep watch on you, and who is responsible for the beady-eyed characters sitting in hydrocars outside. The one who was working for Europol before she married you … before she was instructed to keep a very close eye on a lonely nerd who'd had access to too much dangerous information …" Then he showed me evidence stolen from a Eugov database: the frequent reports Gillian sent to her masters.

I was horrified by the betrayal, but when I got home I said nothing and just watched Gillian carefully. I could not grasp that her smiling manner and loving attentiveness were utterly false, and that I had never been able to see what lay behind them.

MacDoogal was one of the last and most effective political bloggers Europol managed to track down. They sent him to the Milton Keynes indoctrination camps and, like so many sent there, he was never heard from again. On my final visit to Memtech, Jansen revealed that they had cracked another Eugov database and hit the MacDoogal motherlode: hundreds of thousands of private e-mails, psyche and DNA profiles, tens of thousands of images. This, it turned out, was sufficient information to create a recoper: a reconstituted personality.

Read a book, especially non-fiction, and you'll know something about the author. Opinion pieces, as found in blogs, will tell you more. Further detail can be gleaned from the author's responses to others, and from his diaries and from film of him. And much of the organic structure of his brain can be reconstructed from his DNA. Utilizing all of this, Memtech used programs of bewildering complexity, programs that could even make the distinction between irony and sarcasm, to build a model of MacDoogal's functioning mind, then kicked the whole construct into motion in a quantum synaptic computer. He began blogging again, right there on the screen in the Memtech offices, soon tearing into Eugov's every madness.

After Gillian's betrayal I knew I would not long have escaped the camps, and so via a long-prepared secret route, I joined the Memtech staff as they boarded a stealth boat from the Hastings shingle. Some days later when that boat finally slowed beside a jetty in Rabat harbour, I considered how, when reading MacDoogal's blog, one could not know that it was not written by a human being, but then, after my experience with Gillian, who was I to judge façades? ■

Neal Asher is now on his fourth three-book contract with Macmillan, is translated into ten languages, and reckons he's doing OK. His latest, *Line War*, will be out mid-2008.

FUTURES